On well-posedness for partial differential equations. A numerical method for the blunt body problem

by

Per Wahlund

1978

UPPSALA UNIVERSITY

DEPARTMENT OF COMPUTER SCIENCES

On well-posedness for partial differential equations. A numerical method for the blunt body problem

by

Per Wahlund

1978

On well-posedness for partial differential equations. A numerical method for the blunt body problem

Inaugural dissertation

by

Per Wahlund, F. K., Upplands nation

at the University of Uppsala by due permission of the Faculty of Mathematics and the Natural Sciences, to be publicly discussed in room 12, Sturegatan 4B, 2tr, on May 26 1978, at 10 a.m., for the examination of Doctor of Philosophy (according to the Royal proclamation No. 327, 1969).

UNIV.
LEIDEN
BIBL.

The papers summarized here are

A. Wahlund, P.: "Well posedness for mixed initial boundary value problems of general order in one space dimension". Dept. of Comp. Sciences, Uppsala University, Report No 72, 1978.

B. Wahlund, P.: "Finite difference methods for computing the steady flow about blunt bodies at supersonic speed". Dept. of Comp. Sciences, Uppsala University, Report No 73, 1978.

In the papers submitted here we study the solution of partial differential equations, both analytically and numerically and some problems concerning the boundary conditions for these equations. In paper [A] we derive some estimates for the solution of systems of partial differential equations in one space dimension. Both parabolic equations and equations of odd order are considered. In paper [B] a numerical method for solving the equations for steady flow about a blunt body is presented, and different possibilities of specifying the boundary conditions are discussed.

In paper [A] we study the initial boundary value problem for systems of partial differential equations of the type $\frac{\partial u}{\partial t} = \sum_{j=0}^{p} A_j(x) \frac{\partial^j u}{\partial x^j} + f(x,t)$ in the domain $x \geq 0$, $t \geq 0$. The Cauchy problem for the equation on $-\infty < x < \infty$ is assumed to be well posed. We derive algebraic conditions on the highest order terms of the equation and of the boundary conditions at $x = 0$, which enables us to estimate the solution of the full problem. This is done by Laplace-transforming the equations with respect to t and forming a first order system of ordinary differential equations. Estimates of the eigenvalues of the coefficients of this

system are obtained from the theory for the corresponding Cauchy problem. Using perturbation theory we construct a symmetrizer for the full problem, which is defined in a region of the complex plane where the Laplace transform can be inverted, and thus we obtain an estimate of the solution. In the case where the coefficients A_j are functions of x the symmetrizer is constructed from that obtained in the case of constant coefficients by using a partition of unity. The results obtained for parabolic problems are used to derive estimates for the solution of the equations on the domain $0 \leq x \leq 1$, $t \geq 0$ where boundary conditions are given for $x = 0$ and $x = 1$.

In paper [B] we present a finite difference method to compute the shock shape and the steady, inviscid flow about a blunt body at supersonic speed, which computes a solution to the steady flow equations iteratively, without using the time-dependent formulation of the problem. The equations for steady flow are approximated by the same finite difference schemes in both the subsonic and the supersonic region. The difference approximations combined with the boundary conditions give rise to a nonlinear system of equations, which is solved iteratively by Newton's method. The distance of the shock from the body is not known, but is included among the other unknown quantities of the equation, and is adjusted automatically by the iterative procedure as convergence to the solution is obtained. Different types of boundary conditions are discussed with respect to convergence and behaviour of the solution. It is found that for some cases it is favourable to use two different methods successively to obtain a smooth solution, and this procedure have been used for some comparisons with other methods that have been made.

Acknowledgements

I want to express my gratitude to my advisers professor Heinz-Otto Kreiss and docent Bertil Gustafsson for their skilful guidance and encouraging interest in my work.

I would also like to thank all my other friends at the Department of Computer Science for their valuable cooperation in various ways.

DEPARTMENT OF COMPUTER SCIENCE
REPORT No 72
MARCH, 1978

Well posedness for mixed initial boundary value problem of general order in one space dimension

by

Per Wahlund

Notations

If y is a column vector with components y_j, y' denotes its transpose, y^* its adjoint. $u^*y = \sum \overline{u_j}\, y_j$ the Euclidean scalar product of two vectors and $|y| = (y^*y)^{1/2}$ the Euclidean norm of a vector. Similar notations hold for matrices.

For hermitian matrices H_1, H_2 the notation $H_1 > H_2$ means that $y^*H_1y > y^*H_2y$ for all vectors $y \neq 0$. The real part of a matrix is defined by

$$\operatorname{Re} A = (A+A^*)/2$$

The Laplace transform is always used with the dual variable s. The image function is denoted by the capital letter of the original function e.g.

$$U(s) = \int_0^\infty e^{-st}u(t)dt \ .$$

1. Introduction

Consider a system of partial differential equations in one space dimension

$$\frac{\partial u}{\partial t} = P\left(\frac{\partial}{\partial x}\right)u + f(x,t) \tag{1.1}$$

with constant coefficients, where

$$P\left(\frac{\partial}{\partial x}\right) = \sum_{j=0}^{p} A_j \frac{\partial^j}{\partial x^j}$$

is a differential operator of order p with complex $n\times n$ matrices as coefficients and

$$u(x,t) = (u_1(x,t),u_2(x,t),\ldots,u_n(x,t))'$$

is a complex vector function.

We shall always make the following assumptions:

Assumption 1. A_p is nonsingular.

Assumption 2. The imaginary eigenvalues of $(i\omega)^p A_p$ are distinct, ω real.

Assumption 3. The Cauchy problem for (1.1) with $f(x,t) \equiv 0$ is well posed according to the definition given by Kreiss [3]. This means that there are constants a and D so that

$$|e^{(P(i\omega)-aI)t}| \leq D \qquad \omega \text{ real}, \quad t \geq 0 .$$

We want to study the initial boundary value problem for (1.1) in the quarter space $x \geq 0$, $t \geq 0$. We prescribe homogeneous initial values for $t = 0$

$$u(x,0) = 0 \tag{1.2}$$

and boundary conditions for $x = 0$, $t \geq 0$, which are linear relations between u, $\frac{\partial u}{\partial x}$,..., $\frac{\partial^{p-1} u}{\partial x^{p-1}}$. After suitable rearrangement they can be written as a system of linearly independent boundary conditions

$$\sum_{i=0}^{j} L_{ji} \frac{\partial^i}{\partial x^i} u(0,t) = g_j(t) \qquad j=0,1,\ldots,p-1 \tag{1.3}$$

such that the conditions in the principle part of (1.3)

$$L_{jj} \frac{\partial^j}{\partial x^j} u(0,t) = g_j(t) \qquad j=0,1,\ldots,p-1 \tag{1.4}$$

are also linearly independent.

The L_{ji} are constant rectangular $K_j \times n$ matrices, and (1.4) thus represents $K = \sum_{j=0}^{p-1} K_j$ linearly independent boundary conditions. Some of the K_j may be equal to zero.

This and the following problems all have homogeneous initial conditions. However, for the problem above, we will also briefly consider the case with non-homogeneous initial conditions.

As a special case we will consider parabolic systems. The definition of parabolicity is that of Petrovsky [5]:

Definition 1: The system of differential euqations (1.1) is called parabolic if p is even, and there is a constant $\delta > 0$ so that the eigenvalues μ of $(i\omega)^p A_p$ fulfil

$$\operatorname{Re} \mu \leq - \delta |\omega|^p \tag{1.5}$$

for all real ω.

We will here present some special cases, that are considered in this paper.

a) The reduced problem, which consists of the principle part of (1.1)-(1.3)

(1.6) $$\frac{\partial u}{\partial t} = A_p \frac{\partial^p}{\partial x^p} u + f(x,t) \qquad x \geq 0, \quad t \geq 0$$

(1.7) $$u(x,0) = 0 \qquad x \geq 0$$

(1.8) $$L_{jj} \frac{\partial^j}{\partial x^j} u(0,t) = g_j(t) \quad j=0,1,\ldots,p-1 \quad t \geq 0$$

We will give conditions for the reduced problem which enable us to get an estimate for the solution of the full problem (1.1)-(1.3).

b) The problem with coefficients, which are functions of x:

(1.9) $$\frac{\partial u}{\partial t} = \sum_{j=0}^{p} A_j(x) \frac{\partial^j}{\partial x^j} u(x,t) + f(x,t) \quad x \geq 0, \quad t \geq 0$$

(1.10) $$u(x,0) = 0$$

(1.11) $$\sum_{i=0}^{j} L_{ji} \frac{\partial^i}{\partial x^i} u(0,t) = g_j(t) \quad j=0,1,\ldots,p-1 \quad t \geq 0$$

For systems (1.9)-(1.11) of odd order we will make additional assumptions on $A_p(x)$ and $A_{p-1}(x)$.

c) Parabolic problems with two boundaries, which we for convenience take to be on the interval $0 \leq x \leq 1$

(1.12) $$\frac{\partial u}{\partial t} = p(\frac{\partial}{\partial x})u(x,t) + f(x,t) \quad 0 \leq x \leq 1, \quad t \geq 0$$

(1.13) $$u(x,0) = 0 \qquad 0 \leq x \leq 1$$

(1.14) $$\sum_{i=0}^{j} L_{ji} \frac{\partial^i}{\partial x^i} u(0,t) = g_j(t) \quad j=0,1,\ldots,p-1, \quad t \geq 0$$

(1.15) $$\sum_{i=0}^{j} R_{ji} \frac{\partial^i}{\partial x^i} u(1,t) = h_j(t) \quad j=0,1,\ldots,p-1, \quad t \geq 0$$

where the R_{ji} also are rectangular matrices. The results for this problem can also be extended to x-variable coefficients.

For strictly hyperbolic first order systems in several variables the initial boundary value problem has been investigated by Kreiss [4]. We will use basically the same technique for our problems. Eidel'man has in a monograph [1] given a theory for the same problem for parabolic systems. However, his methods can not be generalized to other types of systems.

We shall here summarize the results of this paper. Let $L_2(R)$ denote the space of vector functions which are quadratically integrable over $0 \le x < \infty$, and for $u, v \in L_2(R)$ use the notation

$$(u,v) = \int_0^\infty u^* v \, dx, \quad ||u||^2 = (u,u) .$$

We assume that the function $f(x,t)$ in (1.1) belongs to $L_2(R)$ for all $t \ge 0$ and that $f(x,t)$ and $g_j(t)$, $j=0,1,\ldots,p-1$ are Laplace-transformable with respect to t. For nonparabolic problems we will also assume that for $j=0,1,\ldots,r$, $r = \left[\frac{\tilde{p}}{2}\right] - 1$ the $g_j(t)$ are continuous in t with a piecewise continuous t-derivative and that $g_j(0) = 0$. This means that for $p = 1$ no extra conditions are needed.

The aim is to derive algebraic conditions such that there are constants $K, \alpha > 0$ so that for the solution $u(x,t)$ to (1.1)-(1.3) the estimate

$$(1.16)\quad \int_0^\infty e^{-2\alpha t}(||u(.,t)||^2 + |u(0,t)|^2)dt \le K\int_0^\infty e^{-2\alpha t}(||f(.,t)||^2 + \sum_{j=r+1}^{p-1} |g_j(t)|^2 + \sum_{j=0}^{r} \left|\frac{dg_j(t)}{dt}\right|^2)dt$$

holds.

For parabolic systems we can obtain a sharper estimate

$$(1.17)\quad \int_0^\infty e^{-2\alpha t}(||u(.,t)||^2 + |u(0,t)|^2)dt \le K \int_0^\infty e^{-2\alpha t}(||f(.,t)||^2 + \sum_{j=0}^{p-1} |g_j(t)|^2)d$$

The method is to Laplace transform the system (1.1)-(1.3) with respect to t and obtain a system of ordinary differential equations

$$sU(x,s) = P(\frac{d}{dx})U(x,s) + F(x,s) \tag{1.18}$$

with the boundary conditions

$$\sum_{i=0}^{j} L_{ji} \frac{d^i}{dx^i} U(0,s) = G_j(s) \qquad j=0,1,\ldots,p-1 \tag{1.19}$$

where $U(x,s) = \int_0^\infty e^{-st}u(x,t)dt$.

We write the system (1.18), (1.19) as a first order system by introducing the variables

$$\sigma = s^{\frac{1}{p}} \qquad \text{(principle branch)}$$

$$V_1 = U \qquad \sigma V_j = \frac{dV_{j-1}}{dx} \qquad j=2,\ldots,p$$

$$V = (V_1, V_2, \ldots V_p)'$$

$$F_1 = (0,0,\ldots,A_p^{-1}F)'$$

$$G = (G_0, \sigma^{-1}G_1, \ldots, \sigma^{-(p-1)}G_{p-1})'$$

and obtain

$$\frac{dV}{dx} = \sigma MV - \sigma^{-(p-1)}F_1 \tag{1.20}$$

$$LV(0) = G \tag{1.21}$$

where M and L can be written in the form

$$M = \sum_{j=0}^{p} \sigma^{-j}M_j \tag{1.22}$$

$$L = \sum_{j=0}^{p-1} \sigma^{-j}L_j \ . \tag{1.23}$$

We also transform the reduced problem (1.6)-(1.8) and get

$$\frac{dV}{dx} = \sigma M_0 V - \sigma^{-(p-1)} F_1 \tag{1.24}$$

$$L_0 V(0) = G \tag{1.25}$$

We first consider the Cauchy problem for (1.1) on $-\infty < x < \infty$ with $f(x,t) \equiv 0$ and $u(x,0) = h(x)$, $h \in L_2(-\infty,\infty)$. We assume that the Cauchy problem is well posed according to Kreiss [3]. In section 2 we will derive some estimates for the eigenvalues κ of M, namely

$$|\mathrm{Re}\ \kappa| \geq \delta \frac{\mathrm{Re}\ s}{|s|^{\frac{p-1}{p}} + 1} \qquad \mathrm{Re}\ s \geq 0\ , \tag{1.26}$$

and for parabolic systems the sharper inequality

$$|\mathrm{Re}\ \kappa| \geq \delta |s|^{\frac{1}{p}}\ , \tag{1.27}$$

for $\mathrm{Re}\ s \geq 0$ and $|s|$ sufficiently large.

Let Ω_a denote the region

$$\Omega_a = \left\{ z \in C \;\middle|\; |\mathrm{Arg}\ z| \leq \pi/2p,\ |z| \geq a \right\}$$

and $\partial\Omega_a$

$$\partial\Omega_a = \left\{ z \in C \;\middle|\; |\mathrm{Arg}\ z| = \pi/2p,\ |z| \geq a \right\}$$

part of its boundary. Then

$$\mathrm{Re}\ s \geq 0,\ |s| \geq a^p \iff \sigma \in \Omega_a$$

$$\mathrm{Re}\ s = 0,\ |s| \geq a^p \iff \sigma \in \partial\Omega_a\ .$$

Here a is a suitably chosen constant.

It follows from the inequalities (1.26) and (1.27) that σM has no purely imaginary eigenvalues when σ belongs to the interior of Ω_a and a is sufficiently large. For $\sigma \to \partial\Omega_a$ some of them can approach the imaginary axis. Since the eigenvalues are continuous functions of s, we can determine the number of them with negative real part by letting $\operatorname{Re} s \to \infty$.

We shall show that under certain conditions there exists a transformation $T(\sigma) = \sum_{j=0}^{\infty} \sigma^{-j} T_j$ which is analytical in a neighbourhood of $\sigma = \infty$ and a constant $\delta > 0$ such that

$$T^{-1}(\sigma)(\sigma M)T(\sigma) = \begin{pmatrix} M_{11} & 0 & 0 \\ 0 & M_{22} & 0 \\ 0 & 0 & M_{33} \end{pmatrix} \tag{1.28}$$

where $\operatorname{Re} M_{11} \leq -\delta|\sigma| I$, $\operatorname{Re} M_{33} \geq \delta|\sigma| I$ for $\sigma \in \Omega_a$. M_{22} is a diagonal matrix and some of its eigenvalues can approach the imaginary axis when $\sigma \to \partial\Omega_a$. M_{22} is arranged so that the eigenvalues which have negative real part when σ belongs to the interior of Ω_a are in the first rows. We will also use another parition of $T^{-1}(\sigma M)T$ which is consistent with the first one

$$T^{-1}(\sigma M)T = \begin{pmatrix} M_1 & 0 \\ 0 & M_2 \end{pmatrix} \tag{1.29}$$

where the eigenvalues κ_j of M_j have the property $\operatorname{Re} \kappa_1 < 0$, $\operatorname{Re} \kappa_2 > 0$ for σ in the interior of Ω_a.

The matrix M_0 of the reduced problem (1.24) is obviously transformed to the form (1.29) by T_0.

Consider the eigenvalue problem

$$\begin{cases} \frac{d\phi}{dx} = \sigma M_0 \phi & (1.30) \\ L_0\phi(0) = 0 \qquad \phi \in L_2(R) & (1.31) \end{cases}$$

connected with (1.24), (1.25).

Introduce a new variable $\psi = T_0^{-1}\phi$. Then (1.30), (1.31) becomes

$$\begin{cases} \frac{d\psi}{dx} = \begin{pmatrix} M_{01} & 0 \\ 0 & M_{02} \end{pmatrix}\psi & (1.32) \\ B_0\psi(0) = 0 \qquad \psi \in L_2(R) & (1.33) \end{cases}$$

where (1.33) can be written in partitioned form

$$B_0^{I}\psi^{I}(0) + B_0^{II}\psi^{II}(0) = 0 \ . \qquad (1.33')$$

From the corresponding partitioned form of (1.24), (1.25) we see that in general there will be no solutions if the number of boundary conditions is greater than the number of eigenvalues of M_0 with negative real part, and with not sufficiently many conditions (1.30), (1.31) will have infinitely many solutions. Therefore we have

Lemma 1.1. The problem (1.24), (1.25) has for every G a unique solution $V \in L_2(R)$ only if the number of boundary conditions is euqal to the number of eigenvalues of σM_0 with negative real part for $\sigma \in \Omega_a$, and the eigenvalue problem (1.30), (1.31) has only the trivial solution. The last condition is equivalent to $\text{Det}\,|B_0^{I}| \neq 0$.

The values s with $\text{Re}\, s > 0$ where $\text{Det}\,|B_0^{I}(\sigma)| = 0$ are eigenvalues of the problem (1.30), (1.31). We will define the term generalized

eigenvalue for those s for which (1.30), (1.31) can have eigensolutions which not belong to $L_2(R)$.

Definition 2. A $s = s_0$ with $\operatorname{Re} s_0 = 0$ is called a generalized eigenvalue of (1.30), (1.31) if $\operatorname{Det} |B^I(\sigma_0)| = 0$, $\sigma_0 = s_0^{\frac{1}{p}}$.

We can now formulate the results in a number of theorems. (Assumptions 1 - 3 are assumed to be fulfilled.)

Theorem 1.1. If the problem (1.1)-(1.3) is parabolic, $f(x,t) \in L_2(R)$, $f(x,t)$ and $g_j(t)$, $j=0,1,\ldots,p-1$ are Laplace-transformable with respect to t and the eigenvalue problem (1.30), (1.31) has no eigenvalue or generalized eigenvalue s with $\operatorname{Re} s \geq 0$, there are constants K and $\alpha > 0$ so that the estimate (1.17) holds for the solution $u(x,t)$ of (1.1)-(1.3).

For systems (1.1)-(1.3) of general type we have

Theorem 1.2. If the conditions in Theorem 1.1 are fulfilled and for $j=0,1,\ldots,r = \left[\frac{p}{2}\right] - 1$ the $g_j(t)$ are continuous with piecewise continuous, Laplace-transformable derivative, and $g_j(0) = 0$, there are constants K and $\alpha > 0$ so that the estimate (1.16) holds for the solution $u(x,t)$ of (1.1)-(1.3).

For systems with non-homogeneous initial conditions

$$u(x,0) = h(x) \qquad x \geq 0, \tag{1.34}$$

where we assume $h(x) \in L_2(R)$ and that it is continuous for $x = 0$, we can get an estimate of the solution by writing it as a sum of solutions to problems with homogeneous and non-homogeneous initial conditions.

Theorem 1.3. If the conditions in Theorem 1.1 are fulfilled, but the initial conditions are given by (1.34) an estimate

$$\int_0^\infty e^{-2\alpha t}(||u(.,t)||^2 + |(u(0,t)|^2)dt \leq \tag{1.35}$$

$$K\int_0^\infty e^{-2\alpha t}(||f(.,t)||^2 + \sum_{j=0}^{p-1} |g_j(t)|^2 + ||h||^2 + |h(0)|^2)dt$$

holds.

With additional requirements on $h(x)$ we get for systems of general type:

Theorem 1.4. If the conditions in Theorem 1.2 are fulfilled, but the initial conditions are given by (1.34) where $h(x)$ fulfills $P(\frac{\partial}{\partial x})h(x) \in L_2(R)$ and

$$\sum_{i=0}^{j} L_{ji} \frac{\partial^i}{\partial x^i} h(0) = g_j(0) \qquad j=0,1,\ldots,r = \left[\frac{p}{2}\right] - 1,$$

then an estimate

$$\int_0^\infty e^{-2\alpha t}(||u||^2 + |u(0,t)|^2)dt \leq \tag{1.36}$$

$$K\int_0^\infty e^{-2\alpha t}(||f||^2 + \sum_{j=r+1}^{p-1} |g_j(t)|^2 + \sum_{j=0}^{t} \frac{dg_j(t)}{dt}^2 + ||P(\tfrac{d}{dx})h||^2$$

$$+ \left|\sum_{j=0}^{p-1}\sum_{i=0}^{j} L_{ji} (\tfrac{d}{dx})h(0)\right|^2 + ||h||^2 + |h(0)|^2)dt$$

holds.

By the same technique one can obtain results for problems with variable coefficients. We assume that the coefficients are smooth functions of x and that they converge to constant coefficients when $x \to \infty$, in such a way that there are analytical in a neighbourhood of infinity. The method is to construct a transformation for all x from the transformation in

(1.28) at a number of points $x_0, x_1, \ldots, x_N$ by a partition of unity.

For parabolic problems the result is

Theorem 1.5. If the coefficients $A_j(x)$ in (1.9) are smooth functions of x, converging to constant coefficients when $x \to \infty$, and that $A_p(x)$ fulfil the parabolicity condition in Definition 1. for δ independent of x, and the conditions in Theorem 1.1 are fulfilled, then an estimate of type (1.17) holds for the solution of (1.9)-(1.11).

For systems (1.9)-(1.11) of odd order Assumptions 1, 2 and 3 mean that the eigenvalues of A_p must be real, nonzero and distinct. If we require that this shall hold for $A_p(x)$ uniformly in x we get from perturbation theory that A_p can be transformed to diagonal form with an analytical bounded transformation, and we give the theorem for systems of that form.

Theorem 1.6. If for the system (1.9)-(1.11), where p is odd, the coefficients $A_j(x)$ are smooth functions of x and the following conditions are satisfied uniformly in x:

1. $A_p(x)$ is diagonal, with real, nonzero and distinct eigenvalues.
2. There is a $D > 0$ so that the diagonal elements d_i of $(i\omega)^{p-1}A_{p-1}(x)$ fulfil $\operatorname{Re} d_i < -D|\omega|^{p-1}$ for real ω

$$\left|\frac{d\,A_p(x)}{dx}\right| \Big/ |A_{p-1}(x)| \quad \text{is sufficiently small.}$$

Then an estimate of type (1.17) holds for the solution to (1.9)-(1.11).

Acknowledgements I want to express my gratitude to professor Heinz-Otto Kreiss for introducing this problem to me and for his skilful guidance. I also want to thank docent Bertil Gustafsson and Dr. Björn Engquist for many valueable discussions.

2. The Cauchy problem

Consider the Cauchy problem

$$\frac{\partial v}{\partial t} = P(\frac{\partial}{\partial x})v = \sum_{j=0}^{p} A_j \frac{\partial^j}{\partial x^j} v \qquad -\infty < x < \infty \quad t \geq 0 \tag{2.1}$$

$$v(x,0) = h(x) \qquad h \in L_2(-\infty,\infty) \tag{2.2}$$

and assume that it is well posed in the L_2-sense i.e. there are constants α , $K > 0$ such that the solutions of (2.1) satisfy the estimate

$$||v(x,t)|| \leq K e^{\alpha t}||v(x,0)|| \ . \tag{2.3}$$

Without restriction we can assume that $\alpha \leq 0$. Otherwise we would replace v by $e^{\alpha t}w$.

In [3] it is proved that (2.3) holds with $\alpha \leq 0$ if and only if $(P(i\omega)-sI)^{-1}$ exists for all real ω and all s with $\text{Re } s > 0$ and

$$|(P(i\omega)-sI)^{-1}| \leq K/\text{Re } s \ . \tag{2.4}$$

We shall use (2.4) to prove

Lemma 2.1. Consider the eigenvalue problem

$$P(\kappa)y = sy \qquad \text{Re } s > 0 \ . \tag{2.5}$$

If (2.4) holds then any solution κ of (2.5) satisfies an estimate

$$|\text{Re } \kappa| \geq \frac{\delta|\text{Re } s|}{|s|^{p-1} + 1} \tag{2.6}$$

Here $\delta > 0$ is a constant independent of s.

Proof: By assumption A_p is nonsingular. Therefore there is a constant K_1 such that

$$|\kappa|^p \leq K_1(|s|+1) \tag{2.7}$$

for all solutions of (2.5). Furthermore if $K = \mu + i\omega$, μ,ω real, then

$$\begin{aligned} 0 = \mathrm{Det}|P(\kappa)-sI| &= \mathrm{Det}|P(i\omega+\mu)-sI| = \\ &= \mathrm{Det}|P(i\omega)+\mu Q(i\omega,\mu)-sI| = \\ &= \mathrm{Det}|P(i\omega)-sI|\cdot\mathrm{Det}|I+\mu(P(i\omega)-sI)^{-1}Q(i\omega,\mu)| \end{aligned}$$

$$\mathrm{Det}|P(i\omega)-sI| \neq 0.$$

Therefore the second determinant must vanish. There is also a constant K_2 such that

$$|P(i\omega)-sI)^{-1}Q(i\omega,\mu)| \geq \frac{K_2(|\kappa|^{p-1}+1)}{\mathrm{Re}\ s} \tag{2.8}$$

Therefore

$$|\mu| \geq \frac{1}{K_2}\ \frac{\mathrm{Re}\ s}{|\kappa|^{p-1}+1}$$

and the lemma follows from (2.7).

In the case of parabolic systems the estimate (2.6) can be improved considerably. We have

Lemma 2.2. If the system (2.1) is parabolic then (2.4) becomes

$$|(P(i\omega)-sI)^{-1}| \leq K/|s| \tag{2.9}$$

for s sufficiently large and the estimate (2.6) could be replaced by

$$|\mathrm{Re}\ \kappa| \geq \delta_1|s|^{1/p}\ . \tag{2.10}$$

Proof: We need only consider $|s|$ sufficiently large. We know that for the eigenvalues λ of $(i\omega)^p A_p$ we have $\operatorname{Re}\lambda \leq -\delta|\omega|^p$. For $\omega = 0$ $P(i\omega)$ can be written

$$P(i\omega) = \sum_{j=0}^{p} (i\omega)^j A_j = \omega^p \left\{ \sum_{j=0}^{p} i^j \omega^{j-p} A_j \right\} = \omega^p B(\tfrac{1}{\omega})$$

Since the eigenvalues $\tilde{\lambda}$ of $B(1/\omega)$ are continuous functions of $1/\omega$ we see that they fulfil $\operatorname{Re}\tilde{\lambda} \geq -\delta_2$, $\delta_2 > 0$ for large ω, and hence the eigenvalues of $P(i\omega)$ have real part $\leq -\delta_2 \omega^p$. By [3] $(P(i\omega)-sI)$ can with a bounded transformation be transformed to

$$\begin{pmatrix} \tilde{\lambda}_1\omega^p - s & & b_{12} & & b_{1n} \\ & \tilde{\lambda}_2\omega^p - s & & & \\ & & \ddots & & \\ 0 & & & & \tilde{\lambda}_n\omega^p - s \end{pmatrix}$$

Therefore (2.9) follows, and (2.10) is derived the same way as in Lemma 2.1.

For systems of odd order we can under certain conditions get a similar type of estimate:

Lemma 2.3. Assume that p is odd and that the coefficients in (2.1) fulfil

1. A_p is diagonal with real, distinct eigenvalues,
2. there is a constant $\delta_1 > 0$ so that the diagonal elements d_j of $i^{p-1}A_{p-1}$ fulfil $\operatorname{Re} d_j \geq -\delta_1$.

Then the estimate (2.6) could be replaced by

$$|\operatorname{Re}\kappa| \geq \delta \ . \tag{2.11}$$

Proof: For $\omega \neq 0$ we have

$$P(i\omega) = \omega^p \left\{ \sum_{j=0}^{p} i^j \omega^{j-p} A_j \right\} = \omega^p B(z), \quad z = \frac{1}{\omega} .$$

Since A_p is diagonal with distinct eigenvalues, the eigenvalues of $B(z)$ are analytical at $z = 0$ and their first derivative is given by the corresponding diagonal element of the second term of the sum, i.e. $i^{p-1} A_{p-1}$, see Kato [2]. Therefore there must be constants c_1, $c_2 > 0$ so that the eigenvalues of $P(i\omega)$ for large ω lie to the left of a curve $\tau = \Gamma(\omega)$: $\operatorname{Re} \tau = -c_1 \delta_1 |\omega|^{p-1}$, $\operatorname{Im} \tau = \pm c_2 |\omega|^p$ in the complex plane. Therefore we get $|(P(i\omega)-sI)^{-1}| \leq K/|s|^{\frac{p-1}{p}}$ and (2.11) follows as in Lemma 2.1.

3. The initial boundary value problem

We Laplace transform the system (1.1)-(1.3) with respect to t as in section 1, and write it as the first order system (1.20), (1.21).

The matrices M_j in (1.22) are $p\cdot n \times p\cdot n$ matrices in the form

$$M_0 = \begin{bmatrix} 0 & I & 0 & & 0 \\ 0 & 0 & I & & 0 \\ & & & & \\ 0 & 0 & 0 & & I \\ A_p^{-1} & 0 & 0 & & 0 \end{bmatrix}$$

$$M_k = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ & & & & \\ 0 & 0 & 0 & -A_p^{-1}A_{p-k} \cdots & 0 \end{bmatrix} \qquad k=1,\ldots,p$$

where the nonzero block is in column $(p-k+1)$.

The L_j in (1.23) are $k \times p\cdot n$ matrices where k is the number of boundary conditions

$$L_0 = \begin{bmatrix} L_{00} & & & 0 \\ & L_{11} & & \\ & & L_{22} & \\ & & & \\ 0 & & & L_{p-1,p-1} \end{bmatrix}$$

$$L_k = \begin{bmatrix} 0 & 0 & \cdots & \cdots & 0 \\ \vdots & & & & \\ L_{k0} & & & & \\ & L_{k+1,1} & & & \\ & & & & \\ 0 & & & L_{p-1,p-1-k} \cdots & 0 \end{bmatrix} \qquad k=1,\ldots,p-1$$

By simple calculations we obtain

Lemma 3.1. The eigenvalues of σM are the solutions k of the characteristic equation

$$\text{Det } |P(k)-sI| = 0 .$$

For the transformation of M into block form we have the following lemmas

Lemma 3.2. Assume that $(i\omega)^p A_p$ has the following properties for nonzero and real ω

1. All eigenvalues μ with $\text{Re } \mu = 0$ are distinct
2. The remaining eigenvalues μ satisfy

$$\text{Re } \mu \leq -\delta|\omega|^p$$

for some constant $\delta > 0$.

Then for $\sigma \quad \Omega_a$ there is a transformation T_0 which separates the eigenvalues k of σM_0 into three groups, and constant δ_0, δ_1 so that

$$T_0^{-1}(\sigma M_0)T = \begin{bmatrix} M_{011} & 0 & 0 \\ 0 & M_{022} & 0 \\ 0 & 0 & M_{033} \end{bmatrix}$$

where $(M_{011}+M_{011}^*) \leq -\delta_0|\sigma| I$

$(M_{033}+M_{033}^*) \geq \delta_0|\sigma| I$

and

$$M_{022} = \begin{pmatrix} \kappa_1 & & & 0 \\ & \ddots & & \\ & & \kappa_k & \\ & & & \kappa_{k+1} \\ & & \ddots & \\ 0 & & & \kappa_2 \end{pmatrix}$$

$$\kappa_i \neq \kappa_j \quad i \neq j$$

$$\operatorname{Re} \kappa_j \leq -\frac{\delta_1 \operatorname{Re} s}{|\sigma|^{p-1}} \quad j=1,\ldots,K$$

$$\operatorname{Re} \kappa_j \geq \frac{\delta \operatorname{Re} s}{|\sigma|^{p-1}} \quad j=k+1,\ldots,L$$

Proof: The eigenvalue κ of σM_0 are the $p \cdot n$ values $\kappa_{j\nu} = \lambda_j^{\frac{1}{p}}$, where λ_j are the eigenvalues of sA_p^{-1}. Thus the $\kappa_{j\nu}$ corresponding to the imaginary eigenvalues of $(i\omega)^p A_p$ are all distinct, and by Lemma 2.2 the other eigenvalues are separated in two groups. By elementary transformations we get the inequaliteis for M_{011} and M_{033}. The inequality for M_{022} follows from Lemma 2.1

Lemma 3.3. If A_p fulfils the conditions in Lemma 3.2, then there exists an analytical transformation $T(\sigma) = \sum_{j=0}^{\infty} \sigma^{-j} T_j$ so that

$$T^{-1}(\sigma)(\sigma M)T(\sigma) = \begin{pmatrix} M_{11} & 0 & 0 \\ 0 & M_{22} & 0 \\ 0 & 0 & M_{33} \end{pmatrix}$$

where M_{11}, M_{22}, M_{33} have the same properties as M_{011}, M_{022} and M_{033} provided $\sigma \in \Omega_a$, a sufficiently large. Here T_0 in the expansion is the same as in Lemma 3.2.

Proof: Since $M = \sum_{j=0}^{p} \sigma^{-j} M_j$ is analytical in σ^{-1} and the eigenvalues of M_0 are in three well-separated groups, the lemma follows from perturbation theory, see Kato [2, ch. 2, §1].

We now assume that the conditions of Lemma 1.1 are satisfied. We introduce in (1.20) and (1.21) the variable $W = T^{-1}V$. Note that we here use the partition of $T^{-1}\sigma MT$ defined in (1.29). The equations becomes

$$\begin{cases} \frac{dW}{dx} = \begin{pmatrix} M_1 & 0 \\ 0 & M_2 \end{pmatrix} W - \sigma^{-(p-1)} T^{-1} F_1 & (3.1) \\ B^I W^I(0) + B^{II} W^{II}(0) = G & (3.2) \end{cases}$$

where $B = (B^I \; B^{II}) = L \circ T = \sum_{j=0}^{\infty} \sigma^{-j} B_j$.

Since (1.30), (1.31) has no eigenvalue or generalized eigenvalue for $\mathrm{Re}\, s \geq 0$ we must have $\mathrm{Det}|B_0{}^I| \neq 0$ and since B^I is analytical in σ^{-1} there must be a constant $\beta > 0$ so that $|\mathrm{Det}|B^I|\,| \geq \beta$ for $\sigma \in \Omega_a$, a sufficiently large. We can then express W^I in W^{II} and G at the boundary

$$W^I(0) = -B^{I^{-1}} B^{II} W^{II}(0) + B^{I^{-1}} G \quad . \tag{3.3}$$

Thus there are constants $b_1, b_2 > 0$ so that

$$|W^I(0)|^2 \leq b_1 |W^{II}(0)|^2 + b_2 |G|^2 \quad . \tag{3.4}$$

Let $T(\sigma)$ be the transformation of Lemma 3.3 and consider the hermitian operator

$$S(\sigma) = T(\sigma)^{-1^*} \begin{pmatrix} -cI & 0 \\ 0 & I \end{pmatrix} T(\sigma)^{-1} \tag{3.5}$$

for $\sigma \in \Omega_a$, a sufficiently large. c is a positive constant which will be chosen later. It is clear that S is analytical in σ^{-1} and uniformly bounded in the region.

Let Y be a vector that fulfils the boundary conditions (1.21). Then inequality (3.4) holds for $Z = T^{-1}Y$ and we get

$$Y^*SY = Z^* \begin{pmatrix} -cI & 0 \\ 0 & I \end{pmatrix} Z = -c|Z^I|^2 + |Z^{II}|^2 =$$

$$= c|Z^I|^2 - 2c|Z^I|^2 + |Z^{II}|^2 \geq$$

$$\geq c|Z^I|^2 + (1-2cb_1)|Z^{II}|^2 - 2cb_2|G|^2 \quad .$$

Thus there are constants c_1, $C > 0$ so that

$$Y^*SY \geq c_1|Y|^2 - C|G|^2 \tag{3.6}$$

if we choose c cufficiently small.

We also have

$$\operatorname{Re} S\sigma M = \frac{1}{2}\left(S(\sigma M)+(S(\sigma M))^*\right) =$$

$$= \frac{1}{2} T^{*-1}\left(\left(\begin{pmatrix} -cI & 0 \\ 0 & I \end{pmatrix} T^{-1}\sigma MT + (T^{-1}\sigma MT)^* \begin{pmatrix} -cI & 0 \\ 0 & I \end{pmatrix}\right) T^{-1}\right.$$

$$= \frac{1}{2} T^{*-1}\begin{pmatrix} -c(M_1+M_1^*) & 0 \\ 0 & M_2+M_2^* \end{pmatrix} T^{-1}$$

and by Lemma 3.2 there is a constant c_2 so that

$$\operatorname{Re}(S\sigma M) \geq c_2 \frac{\operatorname{Re} s}{|\sigma|^{p-1}} \quad . \tag{3.7}$$

We can now estimate a solution $V(x)$ of (1.20), (1.21). Since T and T^{-1} are bounded we have $c_1|V| \leq |W| \leq c_2|V|$. From (1.20) we get

$$(3.8)\qquad \mathrm{Re}(V,S\sigma^{-(p-1)}F_1) = \mathrm{Re}\left\{-(V,S\frac{dV}{dx}) + (V,S\sigma MV)\right\} =$$

$$= \mathrm{Re}\left\{\left[-\frac{1}{2}V^*SV\right]_0^\infty + (V,S\sigma MV)\right\} \geq$$

$$\geq \frac{1}{2}c_1|V(0)|^2 - \frac{1}{2}C|G|^2 + c_2\frac{\mathrm{Re}\ s}{|\sigma|^{p-1}}||V||^2$$

and

$$(3.9)\qquad \mathrm{Re}(V,S\sigma^{-(p-1)}F_1) \leq |\sigma|^{-(p-1)}\mathrm{Re}(V,S\begin{pmatrix}0\\0\\\vdots\\A_p^{-1}F\end{pmatrix}) \leq$$

$$\leq K_0|\sigma|^{-(p-1)}(||V||^2+||F||^2)\ .$$

For $\mathrm{Re}\ s$ fixed and suffciently large we get

$$(3.10)\qquad |\sigma|^{-(p-1)}||V||^2 + |V(0)|^2 \leq K_2(|\sigma|^{-(p-1)}||F||^2 + \sum_{j=0}^{p-1}|\sigma|^{-2j}|G_j|^2)$$

or, since U is the first component of V

$$(3.11)\qquad ||U||^2 + |U(0)|^2 \leq K_2(||F||^2 + \sum_{j=0}^{r}|S|^2|G_j|^2 + \sum_{j=r+1}^{p-1}|G_j|^2)$$

where $r = \left[\frac{p}{2}\right] - 1$.

Under the conditions in Theorem 1.2 we can invert the Laplace transform and inequality (1.16) follows from Parseval's relation.

If the system is parabolic we can get a sharper estimate. By Lemma 2.2 the inequality (3.7) can be replaced by

$$(3.12)\qquad \mathrm{Re}(S\sigma M) \geq c_2|\sigma|I$$

and (3.11) becomes

$$(3.13)\qquad ||U||^2 + |U(0)|^2 \leq K_2(|\sigma|^{-(p-1)}||F||^2 + \sum_{j=0}^{p-1}|\sigma|^{-2j}|G_j|^2)$$

and the estimate (1.17) follows.

Thus Theorem 1.1 and Theorem 1.2 are proved.

We shall shortly discuss how to get the estimates in Theorem 1.3 and Theorem 1.4 for equations with non-homogeneous initial conditions. For parabolic systems we write $u(x,t) = v(x,t) + w(x,t)$ where v and w are solutions to

$$(3.14)\qquad \begin{cases} v_t = P(\frac{\partial}{\partial x})v + f \\ v(x,0) = 0 \\ L_j v(0,t) = g_j(t) \qquad j=0,\dots,p-1 \end{cases}$$

and

$$(3.15)\qquad \begin{cases} w_t = P(\frac{\partial}{\partial x})w \\ w(x,0) = h(x) \\ L_j w(0,t) = 0 \qquad j=0,\dots,p-1 \end{cases}$$

where the solution to the Laplace transform of (3.15) can be estimated explicitly. Using Theorem 1.2 on (3.14) and estimating the solution of (3.15) explicitly we obtain the inequality (1.35).

For non-parabolic problems we can, under the conditions in Theorem 1.4, write

$$u(x,t) = v(x,t) + w(x,t) + h(x),$$

where v and w fulfil

$$(3.16)\qquad \begin{cases} v_t = P(\frac{\partial}{\partial x})v + f \\ v(x,0) = 0 \\ L_j v(0,t) = g_j(t) - L_j h(0) \end{cases}$$

and

$$(3.17)\qquad \begin{cases} w_t = P(\frac{\partial}{\partial x})w + P(\frac{\partial}{\partial x})h \\ w(x,0) = 0 \\ L_j w(0,t) = 0 \qquad\qquad j=0,\dots,p-1 \end{cases}$$

and noting that the boundary conditions in (3.16) fulfil the conditions in Theorem 1.2 and that the derivatives $\frac{d}{dt}(g_j - L_j h(0)) = \frac{dg_j}{dt}$ are unchanged, the estimate (1.36) follows from Theorem 1.2.

Systems with variable coefficients

Since first order systems are investigated in [4] we will here assume that $p \geq 2$.

With the assumptions made in Theorem 1.5 and 1.6 we know that for a fix $x = x_0$ the operator $S = S(x_0,\sigma)$ is defined and has the same properties as the operator for the corresponding problem with constant coefficients. The method is to define certain points $x_1,\dots,x_N$ on the x-axis and construct a new operator $S(x,\sigma)$ from $S(x_k,\sigma)$, $k=1,\dots,N$ by a partition of unity. We shall always choose $x_1 = 0$ and $x_N = \infty$, and the new operator $S(x,\sigma)$ is defined

$$(3.18)\qquad S(x,\sigma) = \sum_{k=1}^{N} \phi_k(x) S(x_k,\sigma)$$

where the $\phi_k(x)$ are smooth functions which fulfil

a) $0 \leq \phi_k \leq 1$

b) $\phi_k(x_k) = 1 \qquad k=1,\dots,N-1$

c) $\phi_k(x) = 0 \qquad x \leq x_{k-1},\ x \geq x_{k+1}$

d) $\sum_{k=1}^{N} \phi_k(x) \equiv 1$

e) $\left|\frac{d}{dx}\phi_k(x)\right| \leq \frac{k}{\min(|x_k - x_{k-1}|, |x_k - x_{k+1}|)} \qquad k=1,\dots,N-1$.

The condition b) and e) for $\phi_N(x)$ are

$$\lim_{x\to\infty} \phi_N(x) = 1 \quad \text{and} \quad \left|\frac{d}{dx}\phi_N(x)\right| \leq K .$$

With this definition of $S(x,\sigma)$ we have

a) $S(x,\sigma)$ is uniformly bounded

b) $Y^*S(0,\sigma)Y \geq c_1|Y|^2 - C|G|^2$ for vectors $|Y|$ satisfying the boundary conditions (1.21), (cf. (3.6)),

we also note that

$$\mathrm{Re}(V,S\frac{dV}{dx}) = \frac{1}{2}\left[V^*SV\right]_0^\infty - \frac{1}{2}(V,\frac{dS}{dx}V) .$$

We can now prove Theorem 1.5 for parabolic systems.

For every fixed $x = x_0$ the relation (3.12)

$$\mathrm{Re}(S(x_0,\sigma)\sigma M(x_0)) \geq c_2|\sigma|I$$

holds.

For $|\sigma| \leq$ constant we can, by continuity reasons, to every x_0 find an open interval I_0 containing x_0 so that

(3.19) $$\mathrm{Re}(S(x_0,\sigma)\sigma M(x)) > \frac{c_2}{2}|\sigma|I$$

for all $x \in I_0$.

The conditions on the coefficients ensure that one such interval always can be (d_2,∞) for some d_2. We will also always choose an interval with its left endpoint at $x = 0$. By Heine-Borel's theorem we can cover the rest of the x-axis with a finite number of intervals of the type defined above and we will take the points defining these as our $x_2,\ldots,x_{N-1}$.

For large values of $|\sigma|$ we get, from the expansions (1.22) and $S(x_0,\sigma) = \sum_{j=0}^{\infty} \sigma^{-j} S_j(x_0)$ given by Lemma 3.3, that

$$c_2|\sigma| I \leq \mathrm{Re}(S(x_0,\sigma)\sigma M(x)) =$$

$$= \mathrm{Re}(S_0\sigma M_0) + \mathrm{Re}(S_0M_1+S_1M_0) + \mathcal{O}(\sigma^{-1})$$

and since the dominating term is linear in σ we conclude that we can find a division into intervals of the type above independent of σ.

From the definition (3.18) of $S(x,\sigma)$ we see that it must fulfil (3.19) for all x. The estimate (3.8) now becomes

$$\text{(3.20)} \quad \mathrm{Re}(V,S\sigma^{-(p-1)}F_1) \geq \frac{1}{2} c_1|V(0)|^2 - \frac{1}{2} C|G|^2 - \frac{1}{2}|(V,\frac{dS}{dx} V)| + \frac{c_2}{2}|\sigma| \, ||V||^2$$

and since $|(V,\frac{dS}{dx} V)| \leq K||V||^2$

we can derive the estimate (1.17) as before, if $|\sigma|$ is sufficiently large.

To prove Theorem 1.6 we need

Lemma 3.4. If $A_P(x)$ fulfils the conditions in Theorem 1.6 then there is a real constant $c \geq \frac{1}{2}$ so that

$$\mathrm{Re}(T_0^{-1}(x_0)\sigma M_0(x)T_0(x_0)) = c \, \mathrm{Re}\,(T_0^{-1}(x_0)\sigma M_0(x_0)T_0(x_0))$$

if $|x - x_0|$ is sufficiently small.

T_0 is the first term in the expansion of $T(\sigma)$ in Lemma 3.3.

Proof: The lemma follows from an explicit calculation of $T_0^{-1}(x_0)\sigma M_0(x)T_0(x_0)$.

Write

$$A_p^{-1}(x_0) = \begin{pmatrix} d_1 & & & \\ & \ddots & & \\ & & \ddots & \\ & & & d_n \end{pmatrix}$$

Then the diagonal elements $\tilde{d}_j$ of $A_p^{-1}(x)$ can be written $\tilde{d}_j = (1+\varepsilon_j)d_j$ where the ε_j are real and small if $|x - x_0|$ is sufficiently small.

Then $T_0^{-1}(x_0)\sigma M_0(x)T_0(x_0) =$

$$\begin{pmatrix} D_{11} & D_{12} \cdots & D_{1p} \\ \\ D_{p1} & D_{p2} \cdots & D_{pp} \end{pmatrix}$$

where the D_{ij} are diagonal matrices of order $n \times n$, and

$$D_{ii} = \begin{pmatrix} \omega_{i1}(1+\varepsilon_1/p) & & & \\ & \omega_{i2}(1+\varepsilon_2/p) & & \\ & & \ddots & \\ & & & \omega_{in}(1+\varepsilon_n/p) \end{pmatrix}$$

Here ω_{kj} $k=1,\ldots,p$ are the p values of $\sigma d_j^{1/p}$, i.e. the diagonal elements of $T_0^{-1}(x_0)M_0(x_0)T_0(x_0)$ and the lemma follows.

For the operator S connected with system (1.9)-(1.11) we have from Lemma 2.3

$$\text{Re}(S\sigma M) \geq c \,. \tag{3.21}$$

We shall show that for a given x_0 such an estimate also holds for $\text{Re}(S(x_0,\sigma)\sigma M(x))$ if $|x - x_0|$ is sufficiently small. We can find such

an interval independent of σ if $\sigma = s^{1/p}$ where $s = \xi + i\eta$, $\eta > 0$ fixed and $-\infty < \eta < \infty$. For $|\sigma| \leq$ constant this follows by continuity arguments. For large σ we can write

$$\sigma = |\sigma|(\tau+h(\sigma)|\sigma|^{-p-1})$$

where $\tau = e^{\pm\pi i/2p}$ and $\lim_{\sigma\to\infty} h(\sigma) = \mp i\xi\tau/p$ where the sign depend on whether $\eta \to \infty$ or $\eta \to -\infty$. Using this in the expansion of $\mathrm{Re}(s(x_0,\sigma)\sigma M(x_0))$ we get

$$\begin{aligned}\mathrm{Re}(S(x_0,\sigma)\sigma M(x_0)) &= \mathrm{Re}(S_0\sigma M_0) + \mathrm{Re}(S_0M_1+S_1M_0) + \mathcal{O}(\sigma^{-1}) \\ &= |\sigma|\mathrm{Re}(S_0\tau M_0) + \mathrm{Re}(S_0M_1+S_1M_0) + \mathcal{O}(\sigma^{-1}) \geq c\end{aligned}$$

which gives $\mathrm{Re}(S_0\tau M_0) = 0$.

If $|x - x_0|$ is so small that Lemma 3.4 holds, then we have

$$\begin{aligned}\mathrm{RE}(S(x_0,\sigma)\sigma M(x)) &= \mathrm{Re}(S_0(x_0)\sigma M_0(x)) + \mathrm{Re}(S_0(x_0)M_1(x) + S_1(x_0)M_0(x)) + \mathcal{O}(\sigma^{-1}) = \\ &= c\ \mathrm{Re}(S_0(x_0)\sigma M_0(x_0)) + \mathrm{Re}(S_0(x_0)M_1(x) + S_1(x_0)M_0(x)) + \mathcal{O}(\sigma^{-1}) \\ &= c|\sigma|\mathrm{Re}(S_0(x_0)\tau M_0(x_0)) + \mathrm{Re}(S_0(x_0)M_1(x) + S_1(x_0)M_0(x)) + \mathcal{O}(\sigma^{-1})\end{aligned}$$

and an estimate of type (3.21) follows by continuity arguments. With an operator $S(x,\sigma)$ constructed the same way as for the parabolic system we have an inequality of type (3.8)

$$\begin{aligned}(3.22)\qquad \mathrm{Re}(V,S\sigma^{-(p-1)}F_1) \geq c_1|V(0)|^2 - C|G|^2 - \\ - \tfrac{1}{2}\left|(V,\tfrac{dS}{dx}V)\right| + c_2||V||^2 \ .\end{aligned}$$

Since $(V,\frac{dS}{dx}V) = \sum_{j=0}^{\infty} \sigma^{-j}(V,\frac{dS_j}{dx}V)$

and $S_0(x)$ depends only on $A_p(x)$, we obtain

$$||V||^2 + |V(0)|^2 \leq K(\sigma^{-(p-1)}||F||^2 + \sum_{j=0}^{p-1} |\sigma|^{-2j}|G_j|^2)$$

if $|\sigma|$ is sufficiently large and the derivative $\left| \frac{dA_p(x)}{dx} \right|$ sufficiently small and the estimate (1.17) follows.

Remark: The conditions in Theorem 1.6 are for example fulfilled for equations of type

$$\frac{\partial u}{\partial t} = \varepsilon A_p(x) \frac{\partial^p u}{\partial x^p} + \sum_{j=0}^{p-1} A_j(x) \frac{\partial^j u}{\partial x^j} + f$$

with suitable conditions on A_p and A_{p-1} if $0 < \varepsilon_0 \leq \varepsilon \leq \varepsilon_1$ and ε_1 is sufficiently small.

4. The initial boundary value problem with two boundaries for parabolic systems.

The results for parabolic systems obtained in section 3 can be extended to problems with two boundaries, since the influence of the boundary conditions on each other is small. Consider the system

$$\frac{\partial u}{\partial t} = P(\frac{\partial}{\partial x})u + f(x,t) \qquad 0 \leq x \leq 1, \quad t \geq 0 \tag{4.1}$$

$$u(x,0) = 0 \qquad 0 \leq x \leq 1 \tag{4.2}$$

$$\sum_{i=0}^{j} L_{ji} \frac{\partial^i}{\partial x^i} u(0,t) = g_j(t) \quad j=0,1,\ldots,p-1 \tag{4.3}$$

$$\sum_{i=0}^{j} R_{ji} \frac{\partial^i}{\partial x^i} u(1,t) = h_j(t) \quad t \geq 0 \tag{4.4}$$

With the scalar product and norm defined

$$(u,v) = \int_0^1 u^* v \, dx$$

$$||u||^2 = (u,u)$$

we have

Theorem 4.1. If $P(\frac{\partial}{\partial x})$ is parabolic and the half space problems

$$\frac{\partial u}{\partial t} = P(\frac{\partial}{\partial x})u + f(x,t) \tag{4.5}$$

$$u(x,0) = 0 \qquad 0 \leq x < \infty \tag{4.6}$$

$$\sum_{i=0}^{j} L_{ji} \frac{\partial^i}{\partial x^i} u(0,t) = g_j(t) \tag{4.7}$$

and

$$\frac{\partial u}{\partial t} = P(\frac{\partial}{\partial x})u + f(x,t) \tag{4.8}$$

$$u(x,0) = 0 \qquad -\infty < x \leq 1 \tag{4.9}$$

$$\sum_{i=0}^{j} L_{ji} \frac{\partial^i}{\partial x^i} u(1,t) = h_j(t) \tag{4.10}$$

where $f \equiv 0$ for $x < 0$ and $x > 1$ fulfil the conditions in Theorem 1.1, then there are $\alpha, K > 0$ so that an estimate

$$\int_0^\infty e^{-2\alpha t}(||u||^2+|u(0,t)|^2+|u(1,t)|^2)dt \leq \tag{4.11}$$

$$K \int_0^\infty e^{-2\alpha t}(||f||^2 + \sum_{j=0}^{p-1} (|g_j|^2+|h_j|^2))dt$$

holds for the solution $u(x,t)$ to (4.1)-(4.4).

Proof: We transform the system exactly as was done in section 3. The operator $S(x,\sigma)$ is defined by

$$S(x,\sigma) = T^{*-1}(\sigma)D(x)T^{-1}(\sigma)$$

where

$$D(x) = \begin{pmatrix} (-c-(1-c)x)I & 0 \\ 0 & (1-(1-c)x)I \end{pmatrix}$$

and c is so small that inequalities of type (3.6) with S replaced by $S(0,\sigma)$ and $S(1,\sigma)$ hold at the left and right boundary respectively.

With this definition we have $\frac{dS}{dx} = -(1-c)T^{*^{-1}}T$ and (3.8) becomes

$$\text{(4.12)} \qquad \mathrm{Re}(V,S\sigma^{-(p-1)}F_1) \geq c_1(|V(0)|^2+|V(1)|^2) - c(|G|^2+|H|^2) +$$

$$+ c_2|\sigma|\,||V||^2 - c_3(1-c)||V||^2$$

and for $|\sigma|$ sufficiently large we have same type of estimate as for the parabolic half-space problem and the theorem follows.

Since the operator $S(x,\sigma)$ for large σ has the same properties as the corresponding operator for the half-space problem, one can get the same estimate for problems with variable coefficients under the same conditions as in section 3.

References:

[1] Eidel'man, S. D.: Parabolic systems, North-Holland, Amsterdam 1969.

[2] Kato, T.: Perturbation Theory for Linear Operators, Springer-Verlag, Berlin 1966.

[3] Kreiss, H.-O.: Über sachgemässe Cauchyprobleme, Math. Scand 13 (1963), p. 109-128.

[4] Kreiss, H.-O.: Initial Boundary Value Problems for Hyperbolic Systems, Comm. Pure Appl. Math. 23 (1970), p 277-298.

[5] Petrovsky, I. G.: Lectures on partial differential equations, Interscience, New York 1954.

UPPSALA UNIVERSITY
DEPARTMENT OF COMPUTER SCIENCE
REPORT No 73
APRIL, 1978

Finite Difference Methods for Computing the Steady Flow About Blunt Bodies at Supersonic Speed

by

Per Wahlund

List of symbols

A, B, C, D, E	parameters defined in eq. (2.9)-(2.13)
a	speed of sound
f	vector defined in (2.25)
H	nonlinear function defined in (3.1)
h	enthalphy
K	body curvature parameter
m	number of mesh points in s-direction
n	number of mesh points in y-direction
p	pressure
r	distance from symmetry line
Rn	nose radius of body
s	coordinate along body
Δs	step size in s-direction
t	time
u	velocity tangential to body
v	velocity perpendicular to body
V_∞	free stream velocity
w	vector defined in (2.21)
y	coordinate perpendicular to body
Δy	step size in y-direction
β	angle between tangents to chock and body
γ	ratio of specific heats
δ	shock stand-off distance
θ	angle between tangent to body and symmetry axis
λ, μ	parameters defined in (2.16), (2.17)
ρ	density
ϕ	angle around symmetry axis

1. Introduction

The problem of determining the bow shock shape and the gas flow around a blunt body travelling at supersonic speed, has attained considerable interest for many years, as it has important applications for example in connection with supersonic flight and entry from space into the atmosphere. For this reason many computational methods have been developed for this problem. One possibility is to solve the inverse problem where one from a prescribed shock and conditions across it computes the state of the flow behind it and the body shape, e.g. as in the work of Lomax and Inouye [5]. Another commonly used method is to solve the time-dependent equations governing unsteady flow obtaining the results for the steady flow asymptotically in time. Such a technique is used by Sutton [9] for computing the fully-coupled radiating flow field around a blunt body. The so-called finite volume technique with a split explicit-implicit time differencing is used by Rizzi and Inouye [7] to compute both axisymmetric and fully three-dimensional blunt body flow.

A drawback with the time-dependent techniques is that they are often very timeconsuming since the convergence to the steady state can be slow, and it can also be difficult to determine whether the steady state has been reached or not. We believe that a more efficient way is to solve the equations for steady flow directly. Then one has to solve a problem where the differential equations are of different type in different parts of the region, elliptic in the subsonic region around the stagnation point and hyperbolic in the supersonic region further downstream. In this paper we will present a numerical method for solving the steady state equations for inviscid, isentropic,

axisymmetric and nonrotating flow about a blunt body. This method uses the same finite difference approximations in both the subsonic and supersonic region. This is a technique similar to that used by Blomster and Sköllermo [2], which gave good results for the transonic nozzle flow problem.

The computations are made in a body oriented coordinate system with the shock at a fixed coordinate value. The shock distance δ is treated in the same way as the other unknowns. It also appears in the coefficients of the equations because of the coordinate transformation. The equations and transformations are described in section 2.

The equations are approximated by second order accurate centered difference approximations in the interior of the region of computation At the shock values are given by the Rankine-Hugoniot relations and at the body the rigid wall condition is applied. At the symmetry line and the outflow boundary different combinations of boundary conditions are tested. The difference approximations are described in Section 3 and Appendix C.

The nonlinear system of equations arising from the difference equations is solved iteratively with Newton's method. Gaussian elimination is used to solve the system of linear equations arising at each step at the iterative procedure. In section 4 this is described in more detail.

The results of the numerical tests are presented in Section 5. The results with different combinations of boundary conditions are compared and the effect of using different step sizes in the mesh is studied. In some cases an improved solution is obtained when continuing with one method from the solution of another. Such a technique is used in the cases where comparisons are made with results from other numerical methods.

Acknowledgements

I want to express my gratitude to my adviser docent Bertil Gustafsson for suggesting this problem to me and for many invalueable dicussions during all stages of the project, and to professor Heinz-Otto Kreiss for his encouraging interest in my work.

Thanks are also due to Dr. Gunilla Sköllermo and Jens Blomster for interesting discussions and for the ideas behind the equations solver.

When making the curve plots and diagrams I have been able to use programs developed by Jan Johansson, for which I am very grateful.

I also want to thank Lehna Juteståhl and Karin Molin for their work with the typing of the manuscript.

2. The differential equations.

The nonradiating inviscid flow is governed by the equations of conservation of mass, momentum and energy for unsteady flow. These conservation equations expressed in vector notation are

Mass:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho V) = 0 \tag{2.1}$$

Momentum:

$$\frac{\partial V}{\partial t} + V \cdot \nabla V - \nabla p/\rho = 0 \tag{2.2}$$

Energy:

$$\rho \frac{\partial h}{\partial t} + \rho V \cdot \nabla h - \frac{\partial p}{\partial t} - V \cdot \nabla p = 0, \tag{2.3}$$

where V is the flow velocity vector, ρ the density, p the pressure and h the enthalphy, i.e. energy per unit mass.

We are going to use the same formulation of the equations and the same coordinate systems that is used by Sutton [9] where the shock is assumed to be a smooth surface. The coordinate system is shown in Fig. 1.

The transformations of equations (2.1)-(2.3) to this coordinate system is carried out in [9].

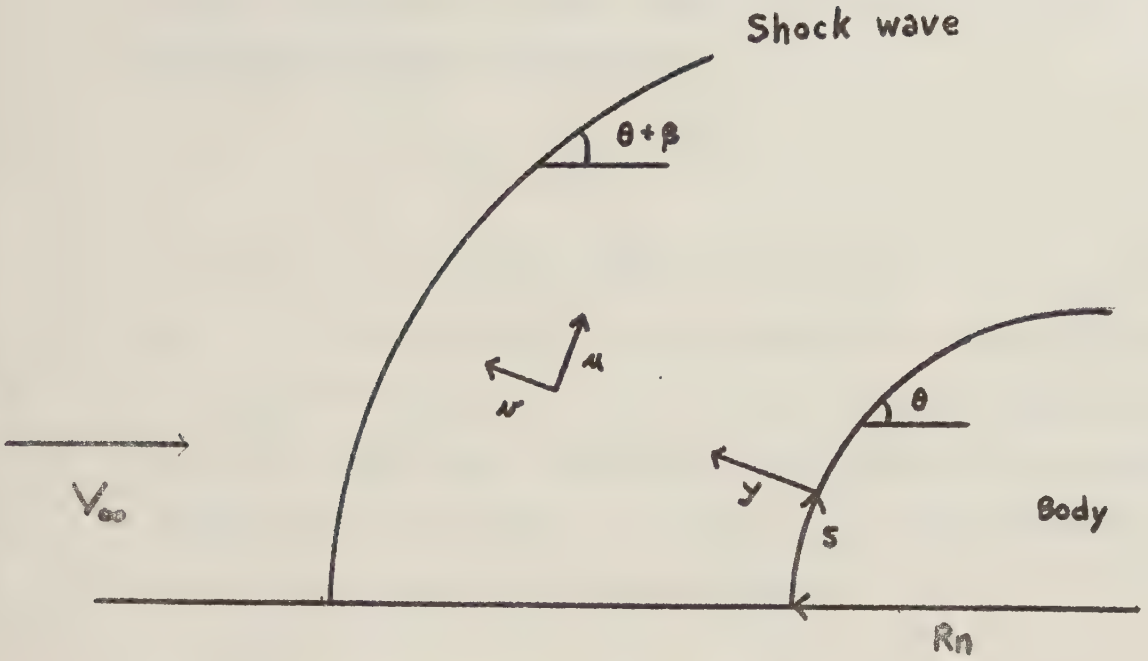


Figure 1: The body-oriented coordinate system.

s	coordinate along body surface
y	coordinate perpendicular to body surface
ϕ	angle measured around symmetry axis (not shown)
θ	angle between tangent to body and symmetry axis
β	angle between tangent to shock wave and symmetry axis
R_n	nose radius of body
u	flow velocity parallel to body
v	flow velocity perpendicular to body

In order to fix the location of the shock wave in the coordinate system we introduce the shock stand-off distance $\delta = \delta(s,t)$ and make the coordinate transformation

$$y' = y/\delta \quad . \tag{2.4}$$

We now have a rectangular coordinate system with $y' = 0$ at the body and $y' = 1$ at the shock. We will hereafter use the unprimed variable y to simplify writing. s is the distance from the symmetry axis along the body surface. With these transformations and assuming axisymmetric flow so that all derivatives and velocities in the ϕ-direction are zero the equations become:

Mass:

$$\rho_t = -(A\rho_y + u\rho_s/\lambda - Bu_y + \rho u_s/\lambda + \rho v_y/\delta + C) \tag{2.5}$$

s-momentum:

$$u_t = -(Au_y + uu_s/\lambda - Bp_y/\rho^2 + p_s/(\rho\lambda) + D) \tag{2.6}$$

y-momentum:

$$v_t = -(Av_y + uv_s/\lambda + p_y/(\delta\rho) - E) \tag{2.7}$$

energy:

$$h_t = -(Ah_y + uh_s/\lambda - p_t/\rho - Ap_y/\rho - up_s/(\rho\lambda)), \tag{2.8}$$

where

$$A = v/\delta - y\delta_t/\delta - uy\mu\tan\beta/(\delta\lambda) \tag{2.9}$$

$$B = \rho y\mu\tan\beta/(\delta\lambda) \tag{2.10}$$

$$C = \rho u \sin\theta/r + \rho v(K/\lambda + \cos\theta/r) \tag{2.11}$$

$$D = Kuv/\lambda \tag{2.12}$$

$$E = Ku^2/\lambda \tag{2.13}$$

Here $$r = y\cos\theta + \int_0^s \sin\theta\, ds \tag{2.14}$$

is the distance from the symmetry axis and

$$K = -\frac{d\theta}{ds} \quad (K=1 \text{ for a sphere}) \tag{2.15}$$

$$\lambda = 1 + K\delta y \tag{2.16}$$

$$\mu = 1 + K\delta \tag{2.17}$$

are variables from the coordinate transformation. All variables appearing in the equations are dimensionless quantities. The relation to dimensional quantities are given in appendix A. Now the pressure p is introduced as independent variable instead of ρ by using the relation

$$\rho = \gamma \cdot p/(\gamma-1) \cdot h \tag{2.18}$$

where γ is the ratio of specific heats for the gas. The mass equation then becomes

$$P_t = -(Ap_y + up_s/\lambda + a^2(\rho u_s/\lambda - Bu_y + \rho v_y/\delta + C)), \tag{2.19}$$

where

$$a = (p\cdot\gamma/\rho)^{1/2} \tag{2.20}$$

a is here the local velocity of sound.

We want the steady state solution, i.e. a solution to the equations (2.6)-(2.8), (2.19) with all time derivatives equal to zero. Writing the independent variables as a vector

$$w = (u\ v\ h\ p)' \tag{2.21}$$

the equations become

$$(w_t) \equiv L(s,y,w)w_s + M(s,y,w)w_y + f(s,y,w) = 0 \tag{2.22}$$

where

$$L = \begin{pmatrix} u/\lambda & 0 & 0 & 1/(\rho\lambda) \\ 0 & u/\lambda & 0 & 0 \\ 0 & 0 & u/\lambda & -u/(\rho\lambda) \\ a^2\rho/\lambda & 0 & 0 & u/\lambda \end{pmatrix} \tag{2.23}$$

$$M = \begin{pmatrix} A & 0 & 0 & -B/\rho^2 \\ 0 & A & 0 & 1/\delta\rho \\ 0 & 0 & A & -A/\rho \\ -a^2B & p_3\rho/\delta & 0 & A \end{pmatrix} \tag{2.24}$$

$$f = \begin{pmatrix} D \\ -E \\ 0 \\ a^2C \end{pmatrix} \tag{2.25}$$

Boundary conditions

We want to solve the equations (2.22) in a region between the body and the shock. Then we have physically very natural boundary conditions at the body and the shock. At the body the normal velocity component v shall be zero. At the shock the values of the variables behind the shock can be calculated from the free stream conditions using the Rankine-Hugouiot relations for a steady shock, when the shock shape is

known, i.e. θ and β are known. The Rankine-Hugouiot relations are given in Appendix B.

The subsonic flow in the stagnation region is followed by a supersonic region furhter downstream along the body. If the region of computation extends sufficiently for downstream in the supersonic region all the characteristics of the time-dependent equations (2.6)-(2.8), (2.19) are pointing out of the domain and no boundary conditions shall be given there. This is also true for the equations (2.22). The values at the outflow boundary are therefore calculated from values upstream using extrapolation. Since the flow is assumed to be axisymmetric it is natural to use the symmetry line as a boundary for the region of computation and using the symmetry of the variable values as boundary conditions. We have also tried the method of specifying the values of the variables at the symmetry axis. Gustafsson and Kreiss [4] has shown that for time-dependent problems such an overspecifying on an artificial boundary leads to slow convergence to the steady state and can also cause oscillations if the given values are not very near the true solution. We have experienced this latter effect also for the steady state problem when making some test calculation. One could of course think of the possibility of solving the one-dimensional steady state problem on the symmetry line to get the proper values. However, the solution of the one-dimensional problem includes an estimate of the tangential derivative $\frac{\partial u}{\partial s}$, and thus too much information is given by the boundary conditions even with this method.

3. The finite difference approximations.

We are going to solve the equations (2.22) using finite-difference approximations in a region between the body and the shock which extends well into the supersonic region. Therefore we choose a value s_{max} and solve the problem on the rectangle $0 \leq s \leq s_{max}$, $0 \leq y \leq 1$. A suitable value of s_{max} can of course not be determined exactly a priori, so we have to choose a reasonable value, and when a solution has been found it should be verified that velocity perpendicular to the boundary really is supersonic. We define a mesh in a rectangle with m points in the s-direction and n points in the y-direction. Hence the meshpoints are (s_j,y_k), $j=1,\ldots,m$, $k=1,\ldots,n$ where $s_j = (j-1) \cdot \Delta s$, $y_k = 1 - (k-1) \cdot \Delta y$ and $\Delta s = s_{max}/(m-1)$ $\Delta y = 1/(n-1)$. Note that the meshpoints in the y-direction are ordered from the shock and inwards; the reason for this is to get a suitable structure of the system of linear equations to be solved in the Newton iterations. For the difference approximations in which symmetry conditions have been used at the symmetry line we have centered the symmetry line halfways between the extreme meshpoints in the s-direction. For such a grid we have $s_j = -\Delta s/2 + (j-1) \cdot \Delta s$ and $\Delta s = s_{max}/(m-1.5)$. For functions $f(s,y)$ of s and y f_{jk} denotes $f(s_j,y_k)$, $f_{j+1/2,k} = (f_{j,k}+f_{j+1,k})/2$ e t c. At each meshpoint we have the values of u, v, h and p as unknowns. The shock distance is also unknown so we have to include the values δ_j, $j=1,\ldots,m$ as unknowns thus making a total of $(4 \cdot m \cdot n+m)$ unknowns.

In the interior of the rectangle we approximate the partial derivatives with second order accurate centered finite differences so that each approximation involves five points. The details of the difference

approximations are described in Appendix C. However, the central difference scheme can not be used for the points on the body, and therefore we use a six-point difference scheme centered halfways between the meshpoints in the y-direction for the points nearest to the body.

At the points on the shock the values of u, v, h and p are given by the Rankine-Hugoniot conditions (see Appendix B), and on the body $v = 0$ is given as boundary condition.

We have made experiments where the values of the variables and the shock distance at the symmetry line are assumed to be known and we have then of course given these values as boundary conditions there. In the case where we use the symmetry of the flow we use the symmetry properties of the variables as boundary conditions; u is antisymmetric and the other variables symmetric. The symmetry of δ is also used as a boundary condition.

Since the values of the variables should not be prescribed at $s = s_{max}$ these values are computed by linear extrapolation from the two adjacent points with the same y-coordinate.

It is also possible to apply an approximation to the differential equation at other points than the mesh points to get the extra conditions that are needed at the outflow boundary. We have tried the four-point box scheme centered halfways between the mesh points in both directions at such conditions. In this case $v_{mn} = 0$ must be specified at the body and no extra condition for δ_m can be used.

The angle β between the tangential directions of the shock and the body is used in the computations of the coefficients in (2.22) We need to compute $\tan\beta$ which is defined by $\tan\beta = \frac{d\delta}{ds} /(1+\delta)$. This quantity is also approximated by central differences where it is possible.

For $s = 0$ we always use $\tan\beta = 0$, but for $s = s_{max}$ we extrapolate it from the values of $\tan\beta$ at adjacent points.

These approximations are second order accurate except for the linear extrapolation at the right boundary, but since no boundary conditions to the differential equations are approximated there we should still be able to get second order accuracy in the interior.

We have also one order lower accuracy in the coefficients one step in from the boundary, due to the extrapolation of δ . Since this is occurs at this s-coordinate only,the over-all accuracy should not be affected.

Description of the difference method used

In the test computations we have worked with several difference methods with different boundary conditions. They all have in common that they use five-point centered difference approximations in the interior, a six-point scheme at the points nearest to the body, values at the shock given by the Rankine-Hugouiot relations and the condition $v = 0$ at the body.

As for the boundary conditions, the combination of symmetry conditions at the symmetry line and linear extrapolation at the outflow boundary has shown to converge for rather general initial approximation.

However, the solution is slightly osccilating in the region near the outflow boundary. A one-dimensional analysis shows that the coefficient for the oscillating term decreases if the box scheme is used at the outflow boundary, and therefore one could expect a smoother solution with that method. However, due to a nearly singular Jacobian we have not been able to get convergence with that method.

Therefore we propose the following procedure:

1. Compute a solution with the method that uses symmetry conditions, and extrapolation at outflow.

2. If oscillations appear in the solution, continue with a method that uses the box scheme at the outflow boundary, and boundary values given on the symmetry line. The symmetry line values and the initial guess for this step should be computed from the solution obtained in the first step.

However, the convergence of the second step seems to depend strongly on the initial guess and the values of the free stream parameters.

We here give the combinations of boundary conditions that have been used, for easy reference later.

Method A Symmetry conditions at symmetry line.
Box scheme used at $s = s_{max}$.

Method B Variable values and shock distance specified at the symmetry line. The box scheme is used at $s = s_{max}$.

Method C Symmetry conditions used at symmetry line. Linear extrapolation used at $s = s_{max}$.

With all unknowns oredered into one vector

$$W = (w_{11}' \; w_{21}' \; \cdots \; w_{m1}' \; w_{12}' \; \cdots \; w_{m2}' \; \cdots \; w_{mn}' \; \delta_1 \; \cdots \; \delta_m)'$$

i.e. first the unknowns at the points on the shock e t c and last the shock distances at the mesh lines, and the equations ordered in a similar manner they give rise to a nonlinear system of equations on the form

$$H(W) \equiv G(W) \circ W + F(W) = 0 \qquad (3.1)$$

Here the equations containing difference quotients have been multiplied by Δy in order to get elements in G of the same order of magnitude.

This system is solved iteratively by Newton's method. All variables, even the δ_j, $j=1,\ldots,m$ are involved in the iterative process, so the adjustment of the shock stand-off distance is made automatically by the Newton procedure. The criterium for convergence is that the norm of the residual, $||H(W)|| = \max_j |H_j(W)|$ should be less than some specified value.

The accuracy in the solution can then be estimated from the correction in the last iteration. The solution method for the system of linear equations is described in section 4.

The initial approximation needed to start the iterative method is obtained in the following way:

1. The shock distance around the body is approximated.
2. The values of u, v, h and p behind the shock is calculated with the Rankine-Hugoniot conditions using free stream values and the approximation of $\tan\beta$ described above.
3. In the interior the values of u, h and p are held constant between the shock and the body, while v is interpolated linearly between the shock-value and the zero value at the body.

The initial approximation of the shock distance used is the one given by Berman in [1] where he computes an analytical expression for the shock distance which involves only the density ratio ρ_∞/ρ across a normal shock.

When the mesh is refined the initial approximation of the unknowns are computed with linear interpolation from the solution on a coarser

mesh. When computations are made using a method with the values on the symmetry line given as boundary conditions, the values can be obtained by interpolating the values from a solution with a method that uses symmetry conditions. This interpolation is done by cubic splines.

4. The linear system of equations.

The Newton iteration procedure used for solving the nonlinear system of equations (3.1) requires the solution of a system of linear equations for every iteration step. The major part of the computing time is spent on these calculations, and the storage requiements for this part tend to grow rapidly when the mesh is refined.

With the oredering of the unknowns and the equations introduced in section 3, the nonlinear system (3.1) has a Jaocbian with the structure that is shown in (4.1). This is a sparse matrix, the main part of which is block tridiagonal, but which also has certain nonzero elements in its last columns.

The reason for this is that the coefficients of the equations are functions of δ and of $\tan\beta$.

$$\left[\begin{array}{cccccc|c} A_1 & B_1 & & & & & d_1 \\ C_2 & A_2 & B_2 & & & & d_2 \\ & \ddots & \ddots & \ddots & & & \\ & & & & B_{n-1} & & \\ & & & C_n & A_n & & d_n \\ \hline & & & & & C_{n+1} & d_{n+1} \end{array}\right] \qquad (4.1)$$

The blocks in the main diagonal in general contain three block diagonals with 4×4 matrices, while the blocks in the super- and sub-diagonal in general have just one such diagonal, but the exact structure depends on which difference approximation that is used. The nonzero portion outside the block diagonals consists of the last m rows and columns of the matrix, where m is the number of mesh points in the s-direction.

This system of equations have been solved by Gaussian elimination, using rowpivoting. The equations were scaled so that the largest element in all rows were of the same magnitude. The scaling factors were always choosen to be a power of the machine radix, (IBM machines have 16 as radix) so that no round-off errors should be caused by the sealing; this type of scaling is recommended by Forsythe and Moler [3].

The method used for the Gaussian elimination is based on algorithms for factorization of band matrices given by Martin and Wilkinson [6], which does not use more storage than is needed for storing the band portion of the matrix plus some extra diagonals needed when pivoting is performed. However, when the mesh is refined storage requirements still grow rapidly. For example, 8 times more storage is needed when the step size in each direction is halved. Therefore we have used the method used by Blomster and Sköllermo in the transsonic nozzle flow problem [2], where only a small part of the band portion of the matrix is kept in main storage at a time, see fig. 2. External storage is then used for the rest of the matrix. This method was modified to suit the special format of the matrix (4.1) and to perform row pivoting. A system with the same coefficient matrix and a new right hand side could be solved rapidly by reading the stored matrix from external storage and performing the eliminations again.

The block diagonals of the system contain many zeroes. For example for the method C described in section 3 the average number of non-zero elements in a row is only 15, regardless of the mesh size. An important factor for the time needed to solve the system is the rate of fill-in, i.e. the number of zero elements that become nonzero during the elimination procedure. There are theories for pivoting strategies

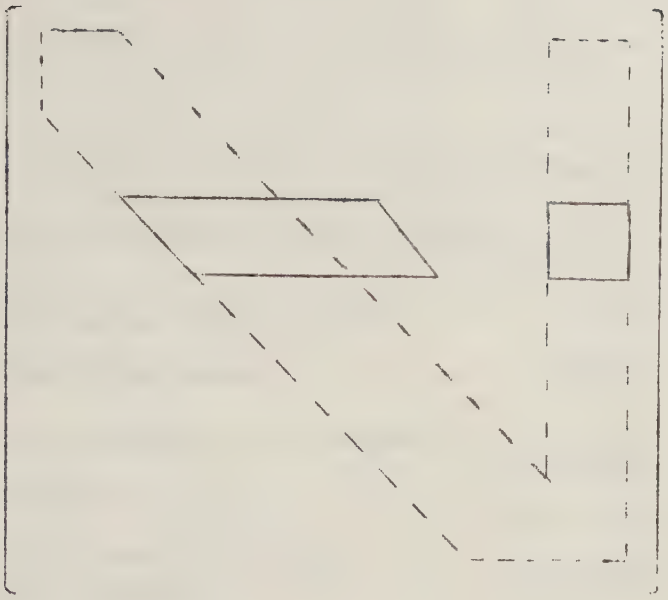


Fig. 2 The part of the matrix held in main storage.

so that the fill-in is held on a minimum or reasonably near it, for example that given by Fewarson [10], but they are based on rather detailed analysis of submatrices of the original matrix at several stages during the elimination process. We have estimated the amount of fill-in during a typical Gaussian elimination by making a count of the number of arithmetical operations needed in the elimination, both when solving the system the way it is normally done, i.e. considering that some elements in the band are zero, and when carrying out all eliminations in the band portion regardless of zeroes within it. For the 7×9 mesh with a total of 259 unknowns, when the matrix has a bound-width of 66 the results are given in Table 1. These figures show that approximately 45% of the elements in the rows and the columns are nonzero at the elimination. This should be compared with the 15% and 20% nonzero elements in those parts of the rows and columns respectively, in the original matrix. The computing time needed to solve the system of equations when the mesh is refined, indicate that the proportion of

in the original matrix. The computing time needed to solve the system

of equations when the mesh is refined,indicate that the proportion of

nonzero elements remains constant or perhaps decreases a little when the number of unknowns becomes larger.

All zeroes within the band considered	240 000
Only zeroes in the columns considered	540 000
No zeroes within the band considered	1 140 000

Table 1. Number of arithmetical operations during Gaussian elimination for 7×9 mesh

Since the matrix (4.1) is very sparse it could be very time-saving if the system of equations could be solved iteratively. However, iterative methods are not likely to be successful for the systems as it is written in (4.1); most of the elements in the main diagonal are small while the elements in the last m columns are fairly large, some of them contain the factor $1/\Delta s$. In the transsonic nozzle flow problem [2] iterative methods for block-tridiagonal systems of similar type were tried, but none of them was as efficient as Gaussian elimination. However, if efficient iterative methods could be developed perhaps more or less specially for systems of the type (4.1) the computing time could be considerably reduced for large meshes.

5. Numerical results.

In order to compare the different methods a test program was written in FORTRAN. The test runs were made on the IBM 370/155 computer at Uppsala University Computing Centre.

The tests were made for the case of a spherical body and the free stream values Mach number $M_\infty = 6.02$ and ratio of specific heat $\gamma = 1.67$. With the nondimensionalizations described in Appendix A we always have the values 1 for the free stream velocity V_∞, density ρ_∞ and the radius R_n of the body. The value for s_{max} was chosen to $3\pi/8 \approx 1.178$ and a mesh with 7 points in the s-direction and 9 points in the y-direction was used, and a refined mesh with 13×17 points.

To get an idea of the structure of the solution a test was first made with the values on the symmetry line given as boundary conditions where different values of the shock stand-off were tried. This does not have any important influence on the shock stand-off at other points, but "wrong" values produced oscillating and physically unrealistic values of u, v, h and p.

Computations with method C on a 7×9 mesh with the initial approximation described in section 3 which uses the shock stand-off estimate given by Berman [2] led to convergence with relatively few iterations. When using Newton's method quadratic convergence should be expected theoretically, but this has not been observed, probably due to that the solution of the linear equations always introduces some rounding errors. We have compared the convergence rate for the pure Newton method with one method that computes a new Jacobian to the system every second iteration and one that uses the same Jacobian for all iterations. The con-

vergence rate for the different strategies is shown in fig. 3. The iterations were continued until the residual $||H(W)||$ was less than 10^{-4}. This condition showed to be unnecessarily restrictive compared to the accuracy of the solution. Instead we have in table 2 shown the number of iterations and computing time needed to make the last correction less than 10^{-3}.

	Number of iterations	Computing time sec.
New Jacobian every iteration	4	43
New Jacobian every second iteration	5	42
Same Jacobian for all iterations	11	61

Table 2: Iterations and computing time to obtain accuracy of 10^{-3} for Method C, 7 × 9 mesh.

We have also computed a solution on a 13 × 17 mesh, where the step-sizes were halved, using the values from the first solution as initial approximations. Convergence was then obtained with 3 iterations, where the Jacobian was recomputed only in the second iteration, giving a computing time of 225 seconds. A comparison of the values obtained for the two meshes is given in table 3. Since the points used by the two meshes are not the same interpolation with cubic splines was used to get comparable values. The error due to interpolation is surely less than 0.3% for the values given there. The shock shape and location of the sonic line obtained with method C is shown in figures 4a - 4b.

The solution obtained with method C shows oscillations near the outflow boundary. These can be the result of a nonaccurate approximation of the variables near the outflow boundary. The extrapolation of δ is the most sensitive as terms $\delta/\Delta s$ occur in the coefficients of the system. Method A avoids this extrapolation but instead gives rise to a nearly singular Jacobian for the Newton method. Therefore method B, which use given values on the symmetry line as boundary condition was used.

The initial approximation and the boundary conditions at the symmetry line were interpolated from the solution obtained with method C and convergence was achieved with 4 - 5 iterations. Comparisons between the solution on a coarse and fine mesh are given in table 4, and show good agreement. Shock shape and diagrams of the shock stand-off distance and pressure distribution are given in figures 5a - 5c.

In the meshes described above the step size in the y-direction is smaller than that in the s-direction, but they still give a sufficiently good resolution in that direction.

The procedure of using method C followed by method B was used when a comparison was made with the inverse method used by Lomax and Inouye [5] for a sphere with the free stream conditions $M_\infty = 5.017$ and $\gamma = 1.4$, see Fig. 6. The main difference to be seen is that the shock distance with our method is about 10% larger, while the density lines show good agreement.

For the case $M_\infty = 10$, $\gamma = 1.4$ we did not get convergence for the second stage of the procedure, the B method. The comparison with the finite volume method by Rizzi and Inouye [7] for this free stream values, shown in fig. 7, was made using method C only.

6. Concluding remarks.

The procedure described in section 3, in which two methods which use different boundary conditions are used in succession, has shown to give reasonable results for the test cases. This procedure includes the computation of a starting approximation, and a solution can therefore be computed with just the free stream conditions given. The computing time needed for the whole procedure is moderate; about 7 minutes is required on an IBM 370/155 to compute a solution using a mesh containing 13 × 17 points. By including the shock distance in the unknowns there is no need to make special adjustments of the shock position, since this is done automatically by the iterative procedure. Neither do we have to determine the location of the sonic line, since the same difference scheme is used all over the domain.

The first stage of this numerical procedure has shown to converge for rather general initial conditions, whereas the second seems sensible to changes in the initial approximation and free stream values. Since the second stage gives a more smooth solution investigations ought to be done to find out what the conditions on the step size, accuracy of the initial approximation etc. should be for that stage to converge. However the convergence of the iterative process is not dependent of the difference approximation only, other methods than the Newton procedure could be more effective in some cases.

We have not been able to extend our mesh up to $s = \pi/2$ without introducing severe oscillations or causing total divergence. The reason for this is probably that the deviation of the mesh from an orthogonal one is relatively large there. The properties of solutions when compu-

tations are made with different types of meshes is investigated by Rizzi [8]. The discussions there concern calculations of flow near Mach number one but the ideas are applicable also to other cases. One could for example think of numerically constructing a mesh where suitable boundary conditions could be given. Then the mesh would be changed as the computations proceed, so that the boundary conditions keep certain desireable properties. If the changes between successive iterations are sufficiently small one would not need to construct a new mesh each step but merely adjust the old one considering the changes of the values of the variables in the latest iteration.

s	0.0	0.5	1.0
Velocity at body	0.0	1.9	0.8
Pressure at body	0.2	2.1	3.8
Shock stand-off distance	1.5	3.2	2.9

Table 3: Percentual difference between result with method C on 7 × 9 and 13 × 17 mesh.

s	0.0	0.5	1.0
Velocity at body	0.0	1.8	1.6
Pressure at body	0.2	0.1	0.8
Shock stand-off distance	1.5	1.0	0.4

Table 4: Percentual difference between result with method B on 7 × 9 and 13 × 17 mesh.

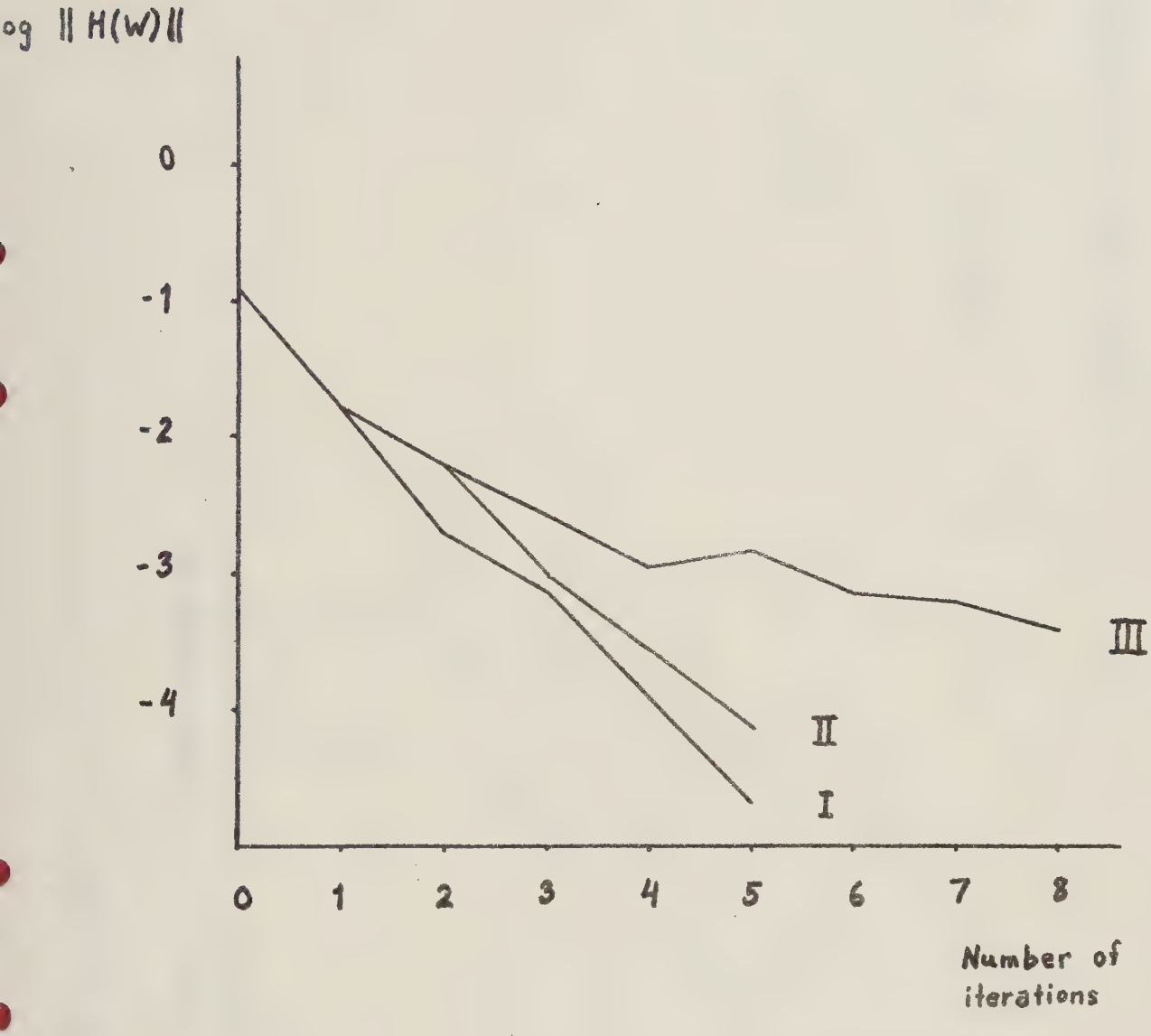


Figure 3: Convergence rate for method C, 7×9 mesh

I New Jacobian computed every iteration

II New Jacobian computed every second iteration

III Same Jacobian used for all iterations
(III converged in 11 iterations)

SHOCK SHAPE AND MACH LINES

METHOD C, 13 X 17 MESH

MACH: 6.02

$\gamma = 1.67$

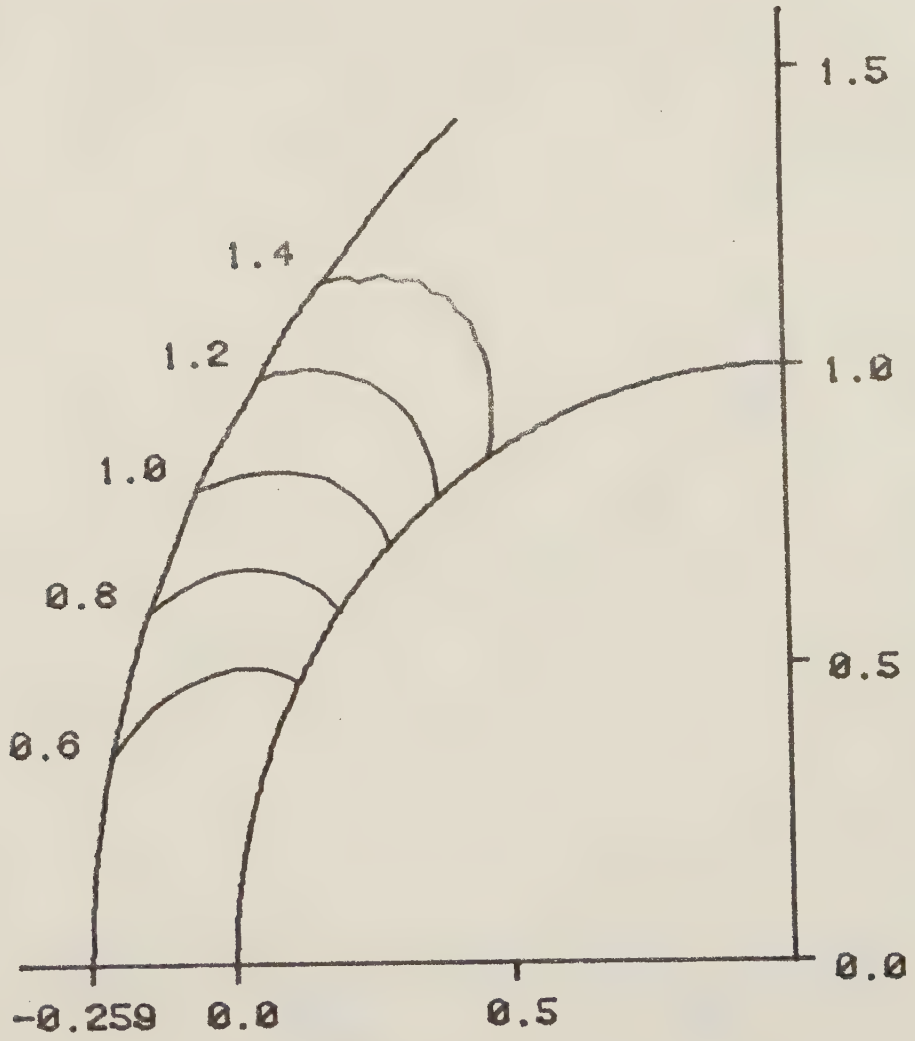
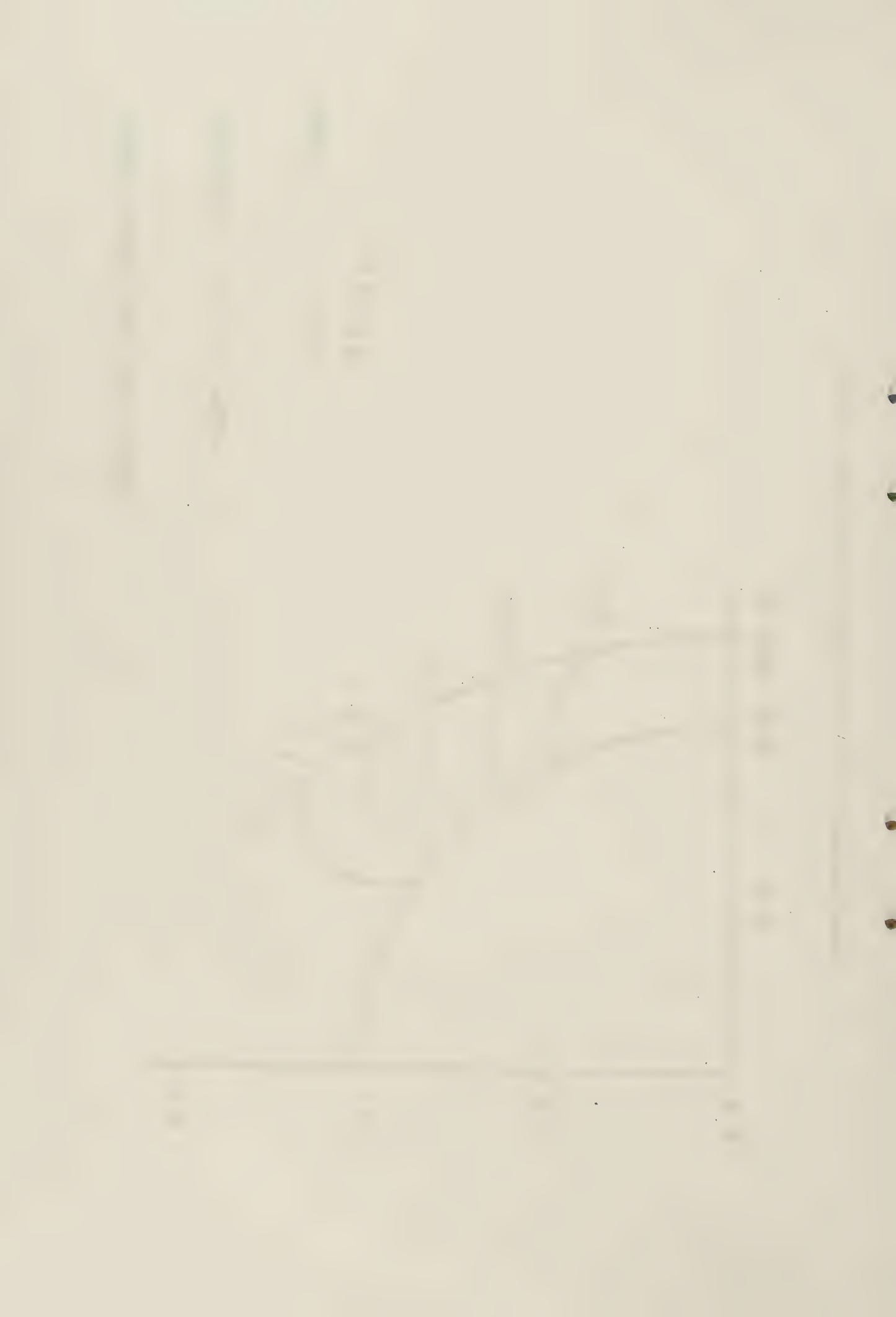


Fig. 48 Shock shape and Mach lines for a spherical body.



SHOCK STAND-OFF DISTANCE

METHOD C, 13 X 17 MESH

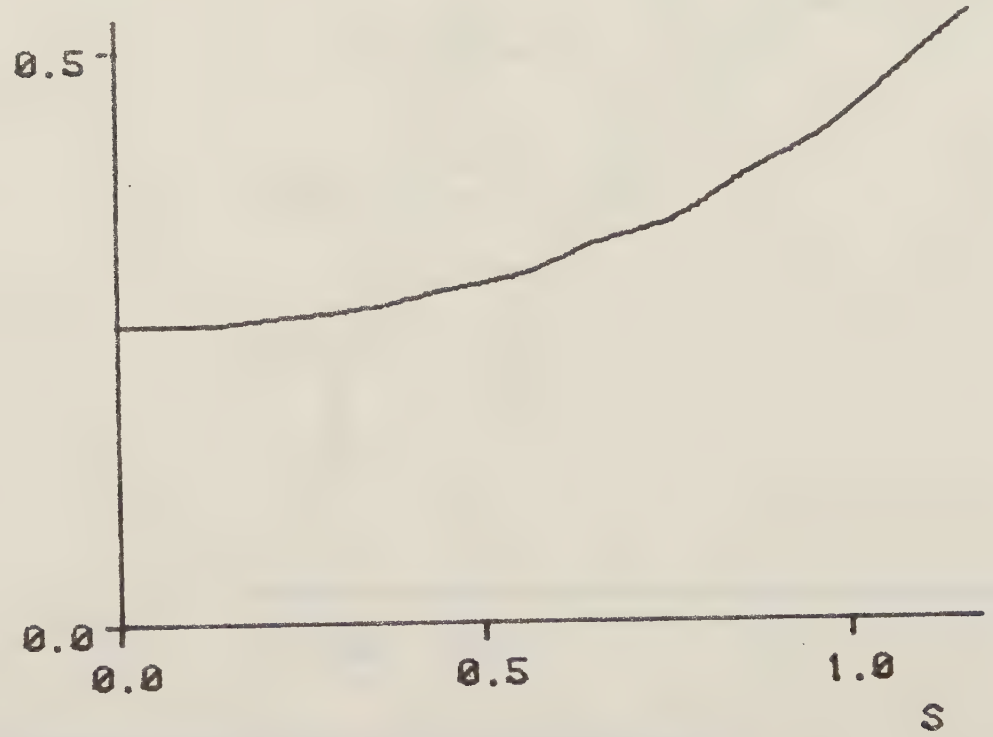


Fig. 4b . Shock stand-off distance for the body in fig. 4a .

SHOCK SHAPE AND MACH LINES

SYMMETRY LINE VALUES FROM METH. C

METHOD B, 13 X 17 MESH

MACH: 6.02

$\gamma = 1.67$

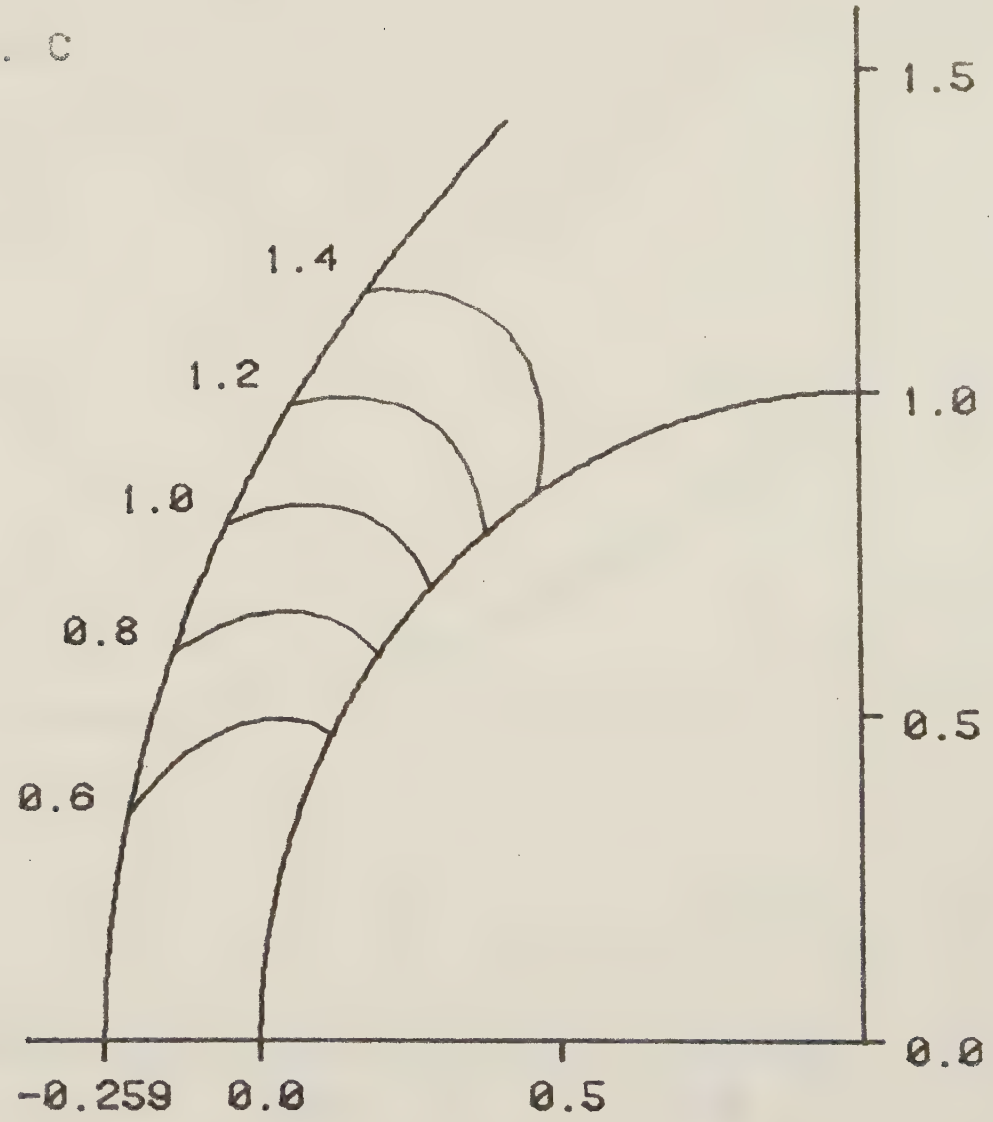
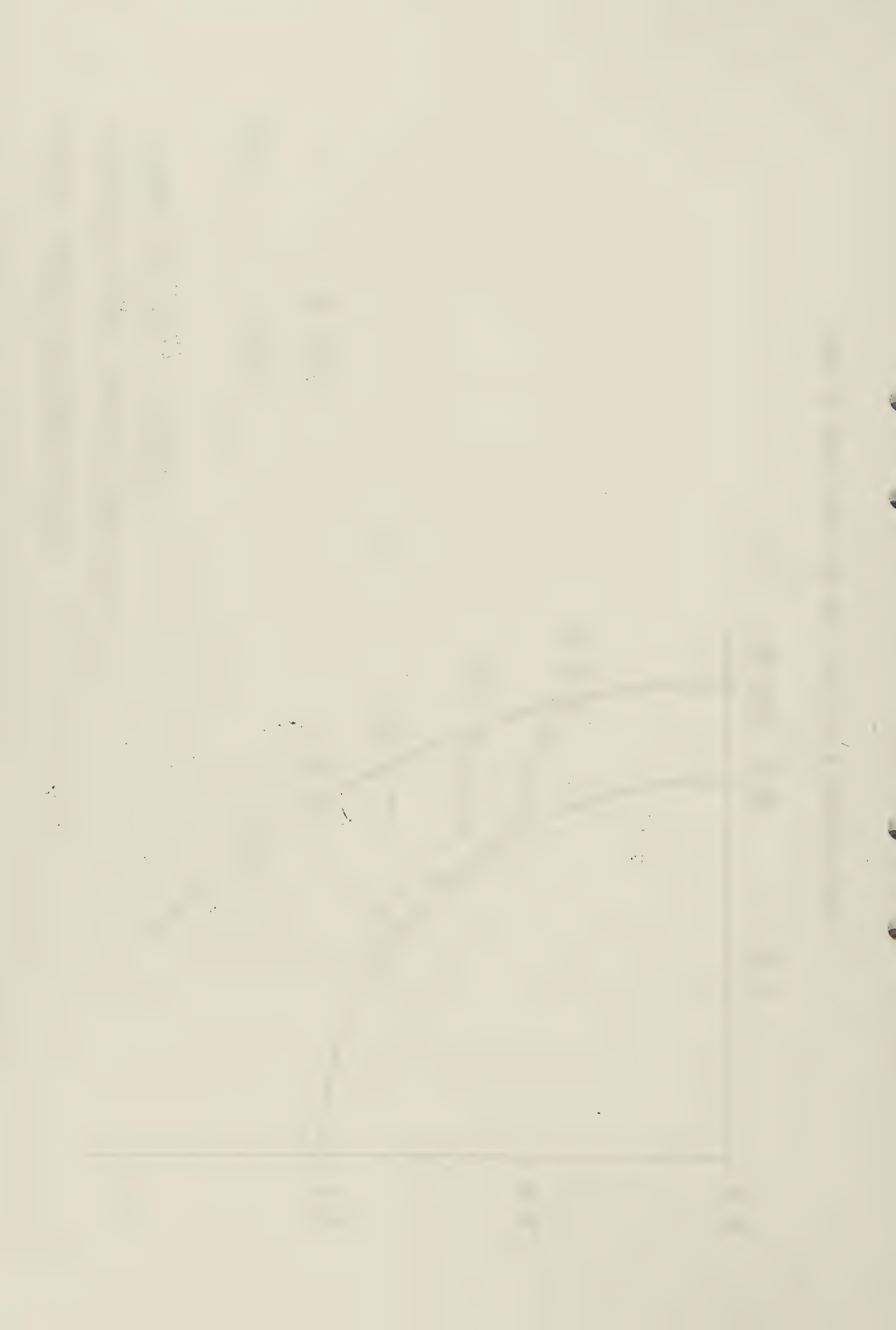


Fig. 5a. Shock shape and Mach lines for a spherical body.



SHOCK STAND-OFF DISTANCE

METHOD B, 13 X 17 MESH

SYMMETRY LINE VALUES FROM METH. C

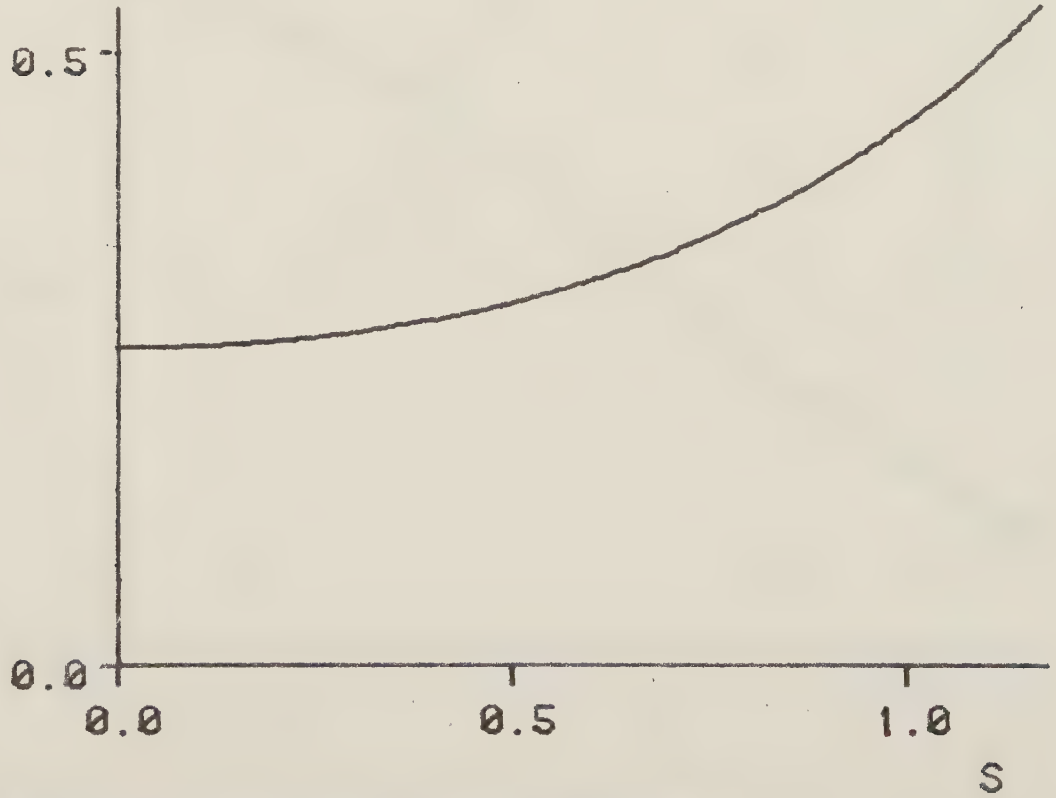
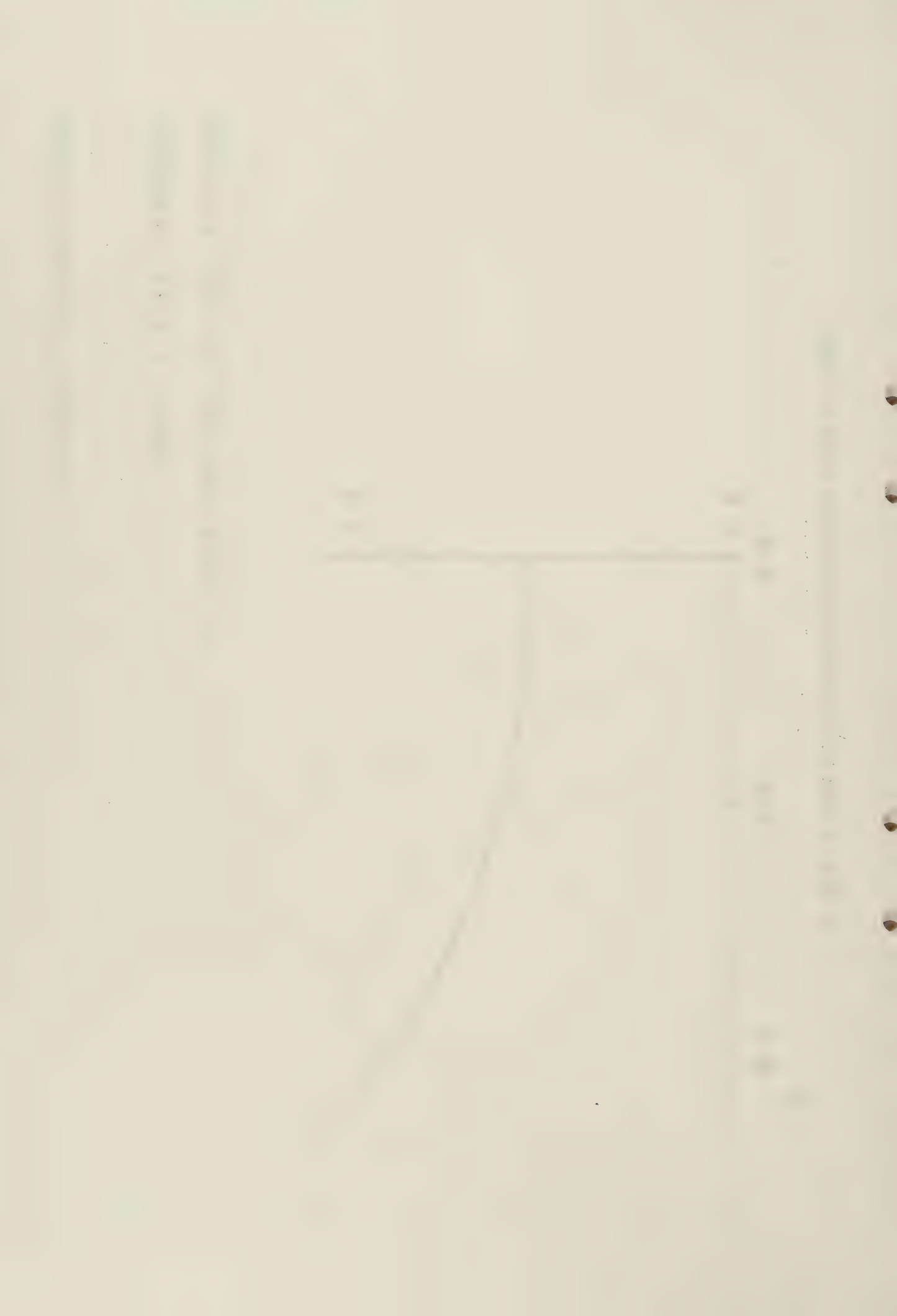


Fig. 5b. Shock stand-off distance for the body in fig. 5a .



PRESSURE DISTRIBUTION ALONG BODY

METHOD B, 13 X 17 MESH

SYMMETRY LINE VALUES FROM METH. C

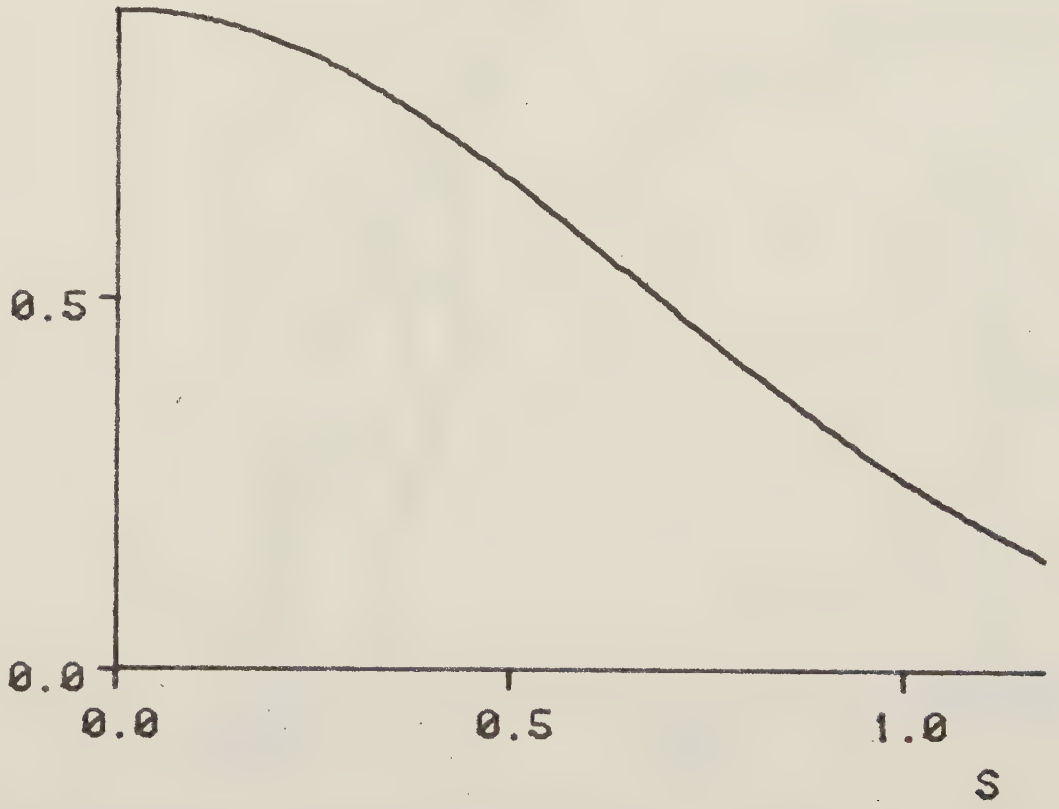


Fig. 5c. Pressure distribution for the body in fig 5a .

SHOCK SHAPE AND DENSITY LINES

PRESENT METHOD B, 13 X 17 MESH ——

LOMAX & INOUYE, (REF. 5) •

MACH: 5.02

$\gamma = 1.4$

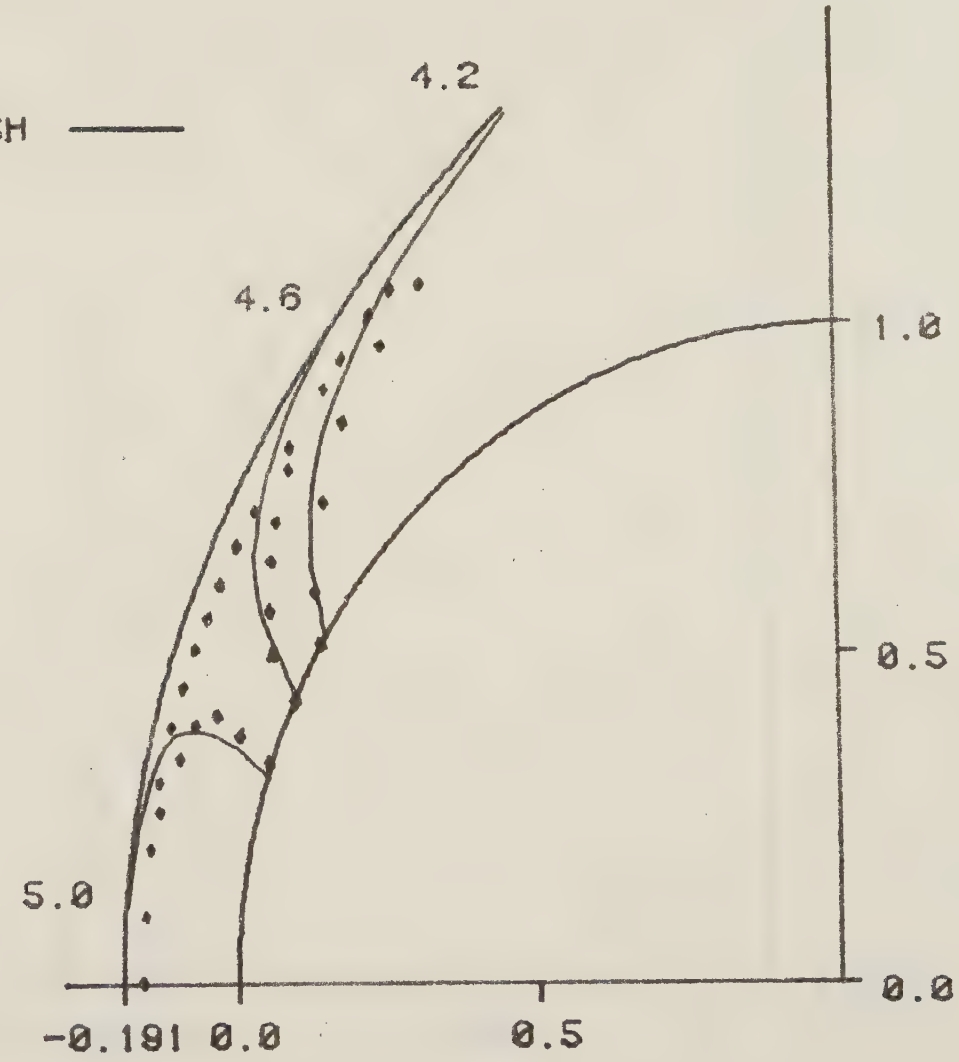


Fig. 6. Shock shape and constant-density lines for a spherical body.

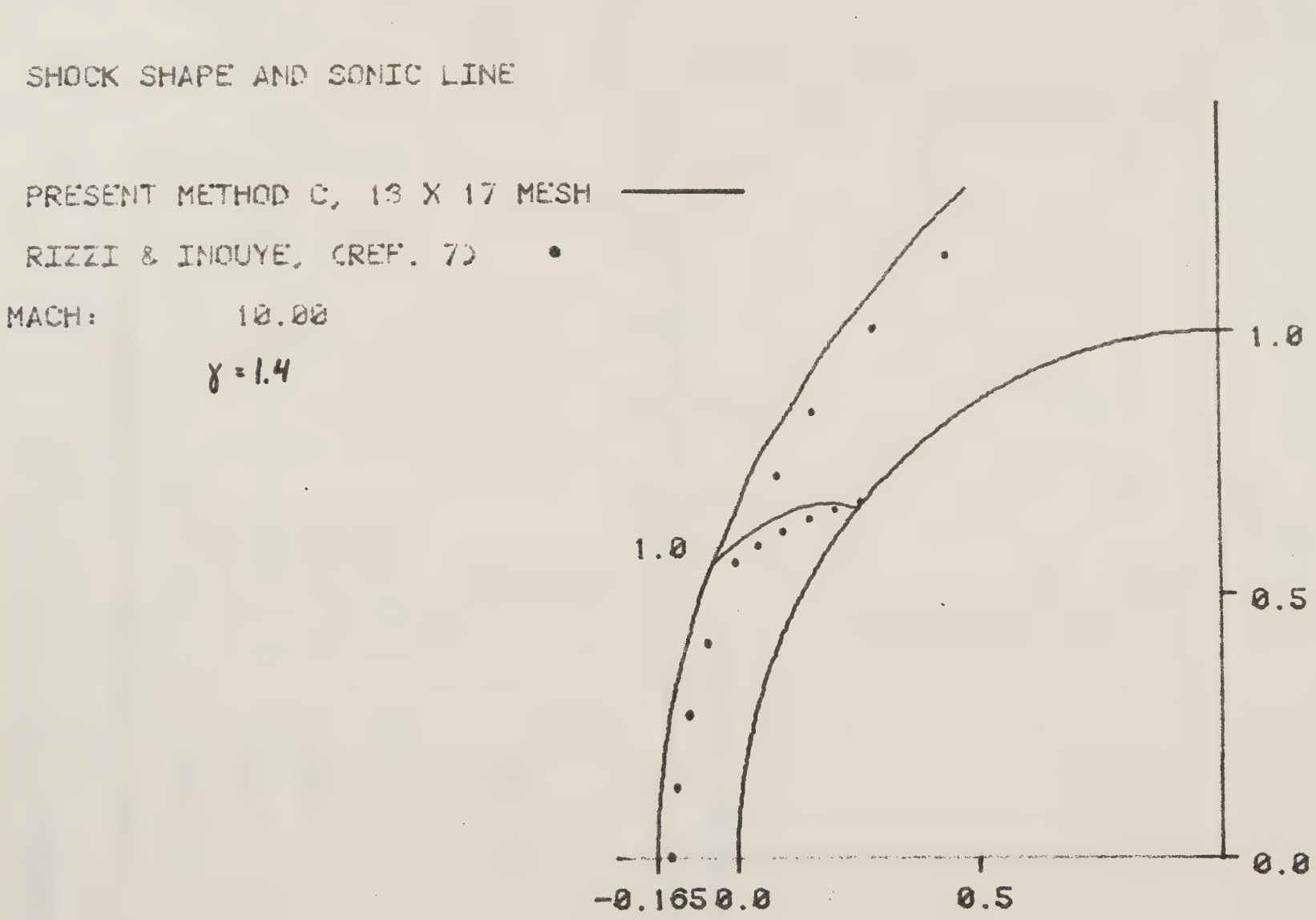


Fig. 7. Shock shape and sonic line for a spherical body.

Appendix A Nondimensionalization

The variables in this report are nondimensionalized by

$$s = s'/Rn'$$

$$y = y'/Rn'$$

$$r = r'/Rn'$$

$$p = p'/(\rho'_\infty V'^2_\infty)$$

$$h = h'/V'^2_\infty$$

$$\rho = \rho'/\rho'_\infty$$

$$t = t' \circ V'_\infty/Rn'$$

where the subscript ∞ denotes free stream values and the prime dimensional quantities.

Appendix B The Rankine-Hugoniot relations in the body coordinate system

The values behind the shock are given by the relations

$$u = V_\infty(\cos(\theta+\beta)\cos\beta + \sin(\theta+\beta)\sin\beta\cdot(2/M_\infty^2+\gamma-1)/(\gamma+1))$$

$$v = V_\infty(\cos(\theta+\beta)\sin\beta - \sin(\theta+\beta)\cos\beta\cdot(2/M_\infty^2+\gamma-1)/(\gamma+1))$$

$$p = p_\infty + 2\rho_\infty V_\infty^2(1-1/M_\infty^2)/(\gamma+1)$$

$$h = \gamma \cdot p\cdot(2/M_\infty^2+\gamma-1)/(\rho_\infty(\gamma-1)(\gamma+1))$$

Appendix C: Description of the difference approximations.

The mesh coordinates are $s_j = s_{min} + (j-1) \cdot \Delta s \quad j=1,\ldots,m$

and $y_k = 1 - (k-1) \cdot \Delta y \quad k=1,\ldots,n$ (note that $y_k > y_{k+1}$)

Functions of s and y are denoted $f_{jk} = f(s_j, y_k)$

and $f_{j+1/2,k} = (f_{j,k} + f_{j+1,k})/2$ etc.

Approximations to the differential equation (2.22):

1. Central difference used in interior:

$$L_{jk}\left(\frac{w_{j+1,k} - w_{j-1,k}}{2\Delta s}\right) + M_{jk}\left(\frac{w_{j,k-1} - w_{j,k+1}}{2\Delta y}\right) + f_{jk} = 0$$

2. Crank-Nicholson scheme used near body:

$$L_{j,n-1/2}\left(\frac{w_{j+1,n-1/2} - w_{j-1,n-1/2}}{2\Delta s}\right) + M_{j,n-1/2}\left(\frac{w_{j,n-1} - w_{j,n}}{\Delta y}\right) + f_{j,n-1/2} = 0$$

3. Box-scheme used near outflow boundary:

$$L_{m-1/2,k+1/2}\left(\frac{w_{m,k+1/2} - w_{m-1,k+1/2}}{\Delta s}\right) + M_{m-1/2,k+1/2}\left(\frac{w_{m-1/2,k-1} - w_{m-1/2,k}}{\Delta y}\right) +$$

$$+ f_{m-1/2,k+1/2} = 0$$

Approximations at boundaries:

1. Symmetry conditions:

$$u_{1,k} = -u_{2,k} \;,\; v_{1,k} = v_{2,k} \;,\; h_{1,k} = h_{2,k} \;,\; p_{1,k} = p_{2,k}$$

$$\delta_1 = \delta_2$$

2. Extrapolation at outflow:

$$w_{m,k} = 2w_{m-1,k} - w_{m-2,k}$$

$$\delta_m = 2\delta_{m-1} - \delta_{m-2}$$

$$\tan \beta_m = 2 \tan \beta_{m-1} - \tan \beta_{m-2}$$

References

1. Berman, R. J.: "Bow shock shape about a spherical nose", AIAA Journ. vol 3 (1965), pp. 778-780.

2. Blomster, J. & Sköllermo G.: "Finite difference computation of steady transsonic nozzle flow", Dept. of Comp. Sciences, Uppsala University, Report no. 66, 1977.

3. Forsythe, G. & Moler, G: "Computer solutions of Linear Algebraic Systems", Prentice-Hall, Englewood Cliffs, 1967.

4. Gustafsson, B. & Kreiss, H.-O.: "Boundary conditions for time dependent problems with an artificial boundary", Dept. of Comp. Sciences, Uppsala University, Report no. 67, 1977.

5. Lomax, H. & Inouye, M.: "Numerical analysis of flow properties about blunt bodies moving at supersonic speed in an equilibrium gas", TR R-204, 1964, NASA.

6. Martin, R. S. & Wilkinson, J. H.: "Solution of symmetric and unsymmetric band equations and the calculation of eigenvectors of band matrices," Handbook for automatic computation, Vol II, pp. 70-92, Springer Verlag, Berlin, Heidelberg, 1971.

7. Rizzi, A. W. & Inouye, M.: "Time-split finite volume method for three-dimensional blunt-body flow", AIAA Journ., vol 11 (1973), pp. 1478-1485.

8. Rizzi, A. W.: "Mesh-related problems in the calculation of supersonic flow near Mach number one", presented at the Conference on numerical methods in applied fluid dynamics, Reading, Jan. 4-6, 1978.

9. Sutton, K.: "Characteristics of coupled nongray radiating gas flow with ablation product effects about blunt bodies during planetary entries", Doctorial thesis at North Carolina State University at Raleigh, 1973.

10. Tewarson, R. P.: "Sparse matrices", Academic Press, New York, 1973.